DUXBURY

Fundamentals of Biostatistics

SIXTH EDITION

Bernard Rosner

Harvard University

THOMSON
BROOKS/COLE

© 2006 Duxbury

Australia • Canada • Mexico • Singapore • Spain
United Kingdom • United States

2 Descriptive Statistics

2.1 Introduction

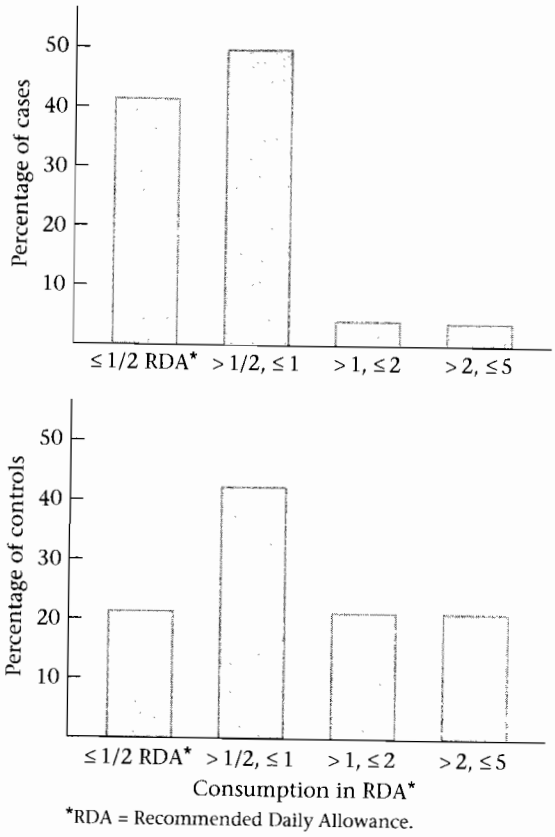
The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

Example 2.1 **Cancer, Nutrition** Some investigators have proposed that consumption of vitamin A prevents cancer. To test this theory, a dietary questionnaire to collect data on vitamin-A consumption among 200 hospitalized cancer cases and 200 controls might be used. The controls would be matched on age and sex to the cancer cases and would be in the hospital at the same time for an unrelated disease. What should be done with these data after they are collected?

Before any formal attempt to answer this question can be made, the vitamin-A consumption among cases and controls must be described. Consider Figure 2.1. The **bar graphs** show visually that the controls have a higher vitamin-A consumption than the cases do, particularly at consumption levels exceeding the Recommended Daily Allowance (RDA).

Example 2.2 **Pulmonary Disease** Medical researchers have often suspected that passive smokers—people who themselves do not smoke but who live or work in an environment where others smoke—might have impaired pulmonary function as a result. In 1980 a research group in San Diego published results indicating that passive smokers did indeed have significantly lower pulmonary function than comparable nonsmokers who did not work in smoky environments [1]. As supporting evidence, the authors measured the carbon-monoxide (CO) concentrations in the working environments of passive smokers and of nonsmokers (where no smoking was permitted in the workplace) to see if the relative CO concentration changed over the course of the day. These results are displayed as a **scatter plot** in Figure 2.2.

Figure 2.1 **Daily vitamin-A consumption among cancer cases and controls**



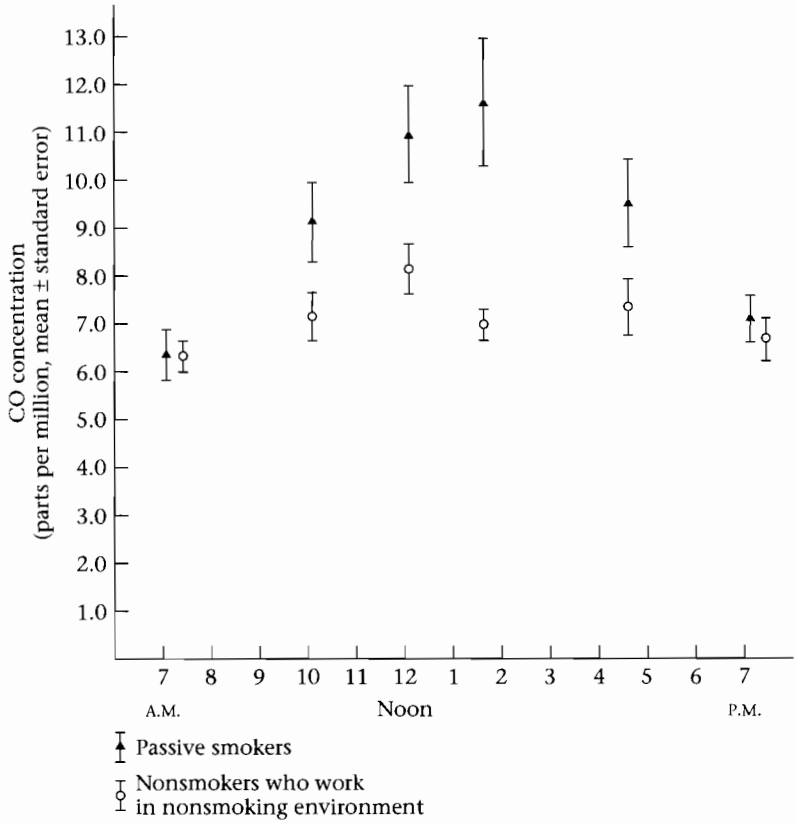
*RDA = Recommended Daily Allowance.

Figure 2.2 clearly shows that the CO concentrations in the two working environments are about the same early in the day but diverge widely in the middle of the day and then converge again after the working day is over at 7 P.M.

Graphic displays illustrate the important role of descriptive statistics, which is to quickly display data to give the researcher a clue as to the principal trends in the data and suggest hints as to where a more detailed look at the data, using the methods of inferential statistics, might be worthwhile. Descriptive statistics are also crucially important in conveying the final results of studies in written publications. Unless it is one of their primary interests, most readers will not have time to critically evaluate the work of others but will be influenced mainly by the descriptive statistics presented.

What makes a good graphic or numeric display? The main guideline is that the material should be as self-contained as possible and should be understandable without reading the text. These attributes require clear labeling. The captions, units, and axes on graphs should be clearly labeled, and the statistical terms used in tables and

Figure 2.2 Mean carbon-monoxide concentration (± standard error) by time of day as measured in the working environment of passive smokers and nonsmokers who work in nonsmoking environments



Source: Reproduced with permission of *The New England Journal of Medicine, 302*, 720–723, 1980.

figures should be well defined. The quantity of material presented is equally important. If bar graphs are constructed, then care must be taken to display neither too many nor too few groups. The same is true of tabular material.

Many methods are available for summarizing data in both numeric and graphic form. In this chapter these methods are summarized and their strengths and weaknesses noted.

2.2 Measures of Location

The basic problem of statistics can be stated as follows: Consider a sample of data $x_1, \ldots, x_n$, where x_1 corresponds to the first sample point and x_n corresponds to the nth sample point. Presuming that the sample is drawn from some population P, what inferences or conclusions can be made about P from the sample?

Before this question can be answered, the data must be summarized as succinctly as possible, because the number of sample points is often large and it is easy to lose track of the overall picture by looking at all the data at once. One type of measure useful for summarizing data defines the center, or middle, of the sample. This type of measure is a **measure of location.**

The Arithmetic Mean

How to define the middle of a sample may seem obvious, but the more you think about it, the less obvious it becomes. Suppose the sample consists of birthweights of all live-born infants born at a private hospital in San Diego, California, during a 1-week period. This sample is shown in Table 2.1.

One measure of location for this sample is the arithmetic mean (colloquially called the *average*). The arithmetic mean (or mean or sample mean) is usually denoted by $\bar{x}$.

Definition 2.1 The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sign Σ (sigma) in Definition 2.1 is a summation sign. The expression

$$\sum_{i=1}^{n} x_i$$

is simply a short way of writing the quantity $(x_1 + x_2 + \cdots + x_n)$.

If a and b are integers, where $a \le b$, then

$$\sum_{i=a}^{b} x_i$$

means $x_a + x_{a+1} + \cdots + x_b$.

Table 2.1 **Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period**

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

If $a = b$, then $\sum_{i=a}^{b} x_i = x_a$. One property of summation signs is that if each term in the summation is a multiple of the same constant c, then c can be factored out from the summation; that is,

$$\sum_{i=1}^{n} cx_i = c\left(\sum_{i=1}^{n} x_i\right)$$

Example 2.3

If $x_1 = 2 \quad x_2 = 5 \quad x_3 = -4$

find $\sum_{i=1}^{3} x_i \quad \sum_{i=2}^{3} x_i \quad \sum_{i=1}^{3} x_i^2 \quad \sum_{i=1}^{3} 2x_i$

Solution

$$\sum_{i=1}^{3} x_i = 2 + 5 - 4 = 3 \qquad \sum_{i=2}^{3} x_i = 5 - 4 = 1$$

$$\sum_{i=1}^{3} x_i^2 = 4 + 25 + 16 = 45 \qquad \sum_{i=1}^{3} 2x_i = 2\sum_{i=1}^{3} x_i = 6$$

It is important to become familiar with summation signs, because they are used extensively in the remainder of the text.

Example 2.4

What is the arithmetic mean for the sample of birthweights in Table 2.1?

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

The arithmetic mean is, in general, a very natural measure of location. One of its main limitations, however, is that it is oversensitive to extreme values. In this instance it may not be representative of the location of the great majority of the sample points. For example, if the first infant in Table 2.1 happened to be a premature infant weighing 500 g rather than 3265 g, then the arithmetic mean of the sample would fall to 3028.7 g. In this instance, 7 of the birthweights would be lower than the arithmetic mean, and 13 would be higher than the arithmetic mean. It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean. The arithmetic mean is a poor measure of central location in these types of samples, because it does not reflect the center of the sample. Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean, is the **median** or, more precisely, the **sample median.**

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

Definition 2.2 The **sample median** is

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The rationale for these definitions is to ensure an equal number of sample points on both sides of the sample median. The median is defined differently when n is even and odd, because it is impossible to achieve this goal with one uniform definition. Samples with an odd sample size have a unique central point; for example, for samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger. Samples with an even sample size have no unique central point, and the middle two values must be averaged. Thus, for samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

Example 2.5 Compute the sample median for the sample in Table 2.1.

Solution First, arrange the sample in ascending order:

2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Because n is even,

Sample median = average of the 10th and 11th largest observations
= (3245 + 3248)/2 = 3246.5 g

Example 2.6 **Infectious Disease** Consider the data set in Table 2.2, which consists of white-blood counts taken on admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

Table 2.2 **Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, PA, on a given day**

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

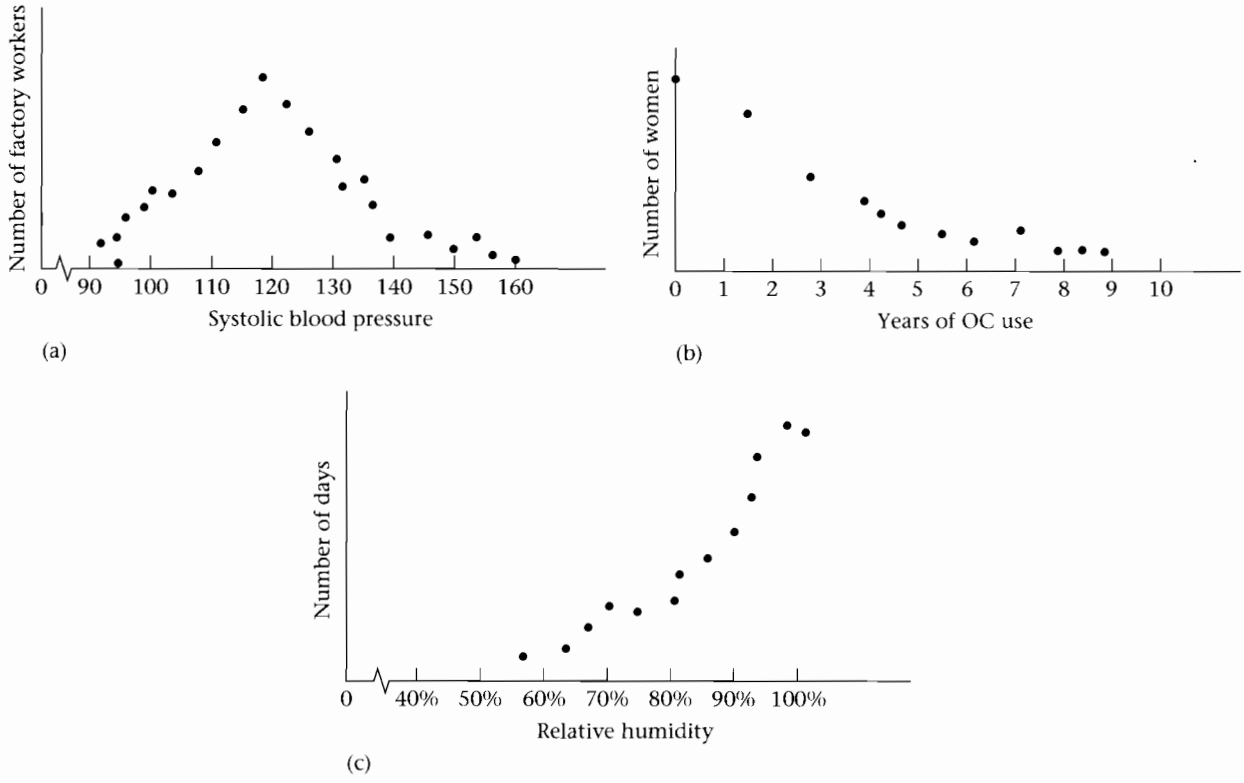
Solution First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35. Because n is odd, the sample median is given by the fifth largest point, which equals 8.

The main strength of the sample median is that it is insensitive to very large or very small values. In particular, if the second patient in Table 2.2 had a white count of 65,000 rather than 35,000, the sample median would remain unchanged, because the fifth largest value is still 8000. Conversely, the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample. The main weakness of the sample median is that it is determined mainly by the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

Comparison of the Arithmetic Mean and the Median

If a distribution is **symmetric,** then the relative position of the points on each side of the sample median are the same. An example of a distribution expected to be roughly symmetric is the distribution of systolic blood-pressure measurements taken on all 30- to 39-year-old factory workers in a given workplace (Figure 2.3a).

Figure 2.3 Graphic displays of (a) symmetric, (b) positively skewed, and (c) negatively skewed distributions


Number of factory workers
0 90 100 110 120 130 140 150 160
Systolic blood pressure
(a)
Number of women
0 1 2 3 4 5 6 7 8 9 10
Years of OC use
(b)
Number of days
0 40% 50% 60% 70% 80% 90% 100%
Relative humidity
(c)

If a distribution is **positively skewed** (skewed to the right), then points above the median tend to be farther from the median in absolute value than points below the median. An example of such a distribution is the distribution of the number of years of oral contraceptive (OC) use among a group of women ages 20–29 years (Figure 2.3b). Similarly, if a distribution is **negatively skewed** (skewed to the left), then points below the median tend to be farther from the median in absolute value than points above the median. An example of such a distribution is the distribution of relative humidities observed in a humid climate at the same time of day over a number of days. In this case, most humidities are at or close to 100%, with a few very low humidities on dry days (Figure 2.3c).

In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution. In particular, for symmetric distributions the arithmetic mean is approximately the same as the median. For positively skewed distributions, the arithmetic mean tends to be larger than the median; for negatively skewed distributions, the arithmetic mean tends to be smaller than the median.

The Mode

Another widely used measure of location is the mode.

Definition 2.3 The **mode** is the most frequently occurring value among all the observations in a sample.

Example 2.7 **Gynecology** Consider the sample of time intervals between successive menstrual periods for a group of 500 college women aged 18–21, shown in Table 2.3. The frequency column gives the number of women who reported each of the respective durations. The mode is 28, because it is the most frequently occurring value.

Table 2.3 **Sample of time intervals between successive menstrual periods (days) of college-age women**

Value	Frequency	Value	Frequency	Value	Frequency
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

Example 2.8 Compute the mode of the distribution in Table 2.2.

Solution The mode is 8000, because it occurs more frequently than any other white-blood count.

Some distributions have more than one mode. In fact, one useful method of classifying distributions is by the number of modes present. A distribution with one mode is called **unimodal**; two modes, **bimodal**; three modes, **trimodal**; and so forth.

Example 2.9 Compute the mode of the distribution in Table 2.1.

Solution There is no mode, because all the values occur exactly once.

Example 2.9 illustrates a common problem with the mode: It is not a useful measure of location if there is a large number of possible values, each of which occurs infrequently. In such cases the mode will be either far from the center of the sample or, in extreme cases, will not exist, as in Example 2.9. The mode is not used in this text because its mathematical properties are, in general, rather intractable, and in most common situations it is inferior to the arithmetic mean.

The Geometric Mean

Many types of laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques, can be expressed either as multiples of 2 or as a constant multiplied by a power of 2; that is, outcomes can only be of the form $2^k c$, $k = 0, 1, \ldots$, for some constant c. For example, the data in Table 2.4 represent the minimal inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients [2]. The arithmetic mean is not appropriate as a measure of location in this situation, because the distribution is very skewed.

However, the data do have a certain pattern, because the only possible values are of the form $2^k(0.03125)$ for $k = 0, 1, 2, \ldots$. One solution is to work with the distribution of the logs of the concentrations. The log concentrations have the property that successive possible concentrations differ by a constant; that is, $\log(2^{k+1}c) - \log(2^k c) = \log(2^{k+1}) + \log c - \log(2^k) - \log c = (k + 1) \log 2 - k \log 2 = \log 2$. Thus, the log concentrations are equally spaced from each other, and the resulting distribution is now not as skewed as the concentrations themselves. The arithmetic mean can then be computed in the log scale; that is,

Table 2.4 Distribution of minimal inhibitory concentration (MIC) of penicillin G for *N. gonorrhoeae*

Concentration (μg/mL)	Frequency	Concentration (μg/mL)	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

Source: Reproduced with permission from *JAMA, 220,* 205–208, 1972. Copyright 1972, American Medical Association.

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

and used as a measure of location. However, it is usually preferable to work in the original scale by taking the antilogarithm of $\overline{\log x}$ to form the geometric mean, which leads to the following definition:

Definition 2.4 The **geometric mean** is the antilogarithm of $\overline{\log x}$, where

$$\overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

Any base can be used to compute logarithms for the geometric mean. The geometric mean is the same regardless of which base is used. The only requirement is that the logs and antilogs in Definition 2.4 should be in the same base. Bases often used in practice are base 10 and base e; logs and antilogs using these bases can be computed using many pocket calculators.

Example 2.10 **Infectious Disease** Compute the geometric mean for the sample in Table 2.4.

Solution (1) For convenience, use base 10 to compute the logs and antilogs in this example.

(2) Compute

$$\overline{\log x} = \begin{bmatrix} 21\log(0.03125) + 6\log(0.0625) + 8\log(0.125) \\ +\ 19\log(0.250) + 17\log(0.50) + 3\log(1.0) \end{bmatrix} \Big/ 74 = -0.846$$

(3) The geometric mean = the antilogarithm of $-0.846 = 10^{-0.846} = 0.143$.

2.3 Some Properties of the Arithmetic Mean

Consider a sample $x_1, \ldots, x_n$, which will be referred to as the original sample. To create a **translated sample** $x_1 + c, \ldots, x_n + c$, add a constant c to each data point. Let $y_i = x_i + c$, $i = 1, \ldots, n$. Suppose we want to compute the arithmetic mean of the translated sample. We can show that the following relationship holds:

Equation 2.1

If $y_i = x_i + c, \quad i = 1, \ldots, n$

then $\bar{y} = \bar{x} + c$

Therefore, to find the arithmetic mean of the y's, compute the arithmetic mean of the x's and add the constant c.

This principle is useful because it is sometimes convenient to change the "origin" of the sample data—that is, to compute the arithmetic mean after the translation and transform back to the original origin.

Example 2.11 In Table 2.3 it is more convenient to work with numbers that are near zero than with numbers near 28 to compute the arithmetic mean of the time interval between menstrual periods. Thus a translated sample might first be created by subtracting 28 days from each outcome in Table 2.3. The arithmetic mean of the translated sample could then be found and 28 added to get the actual arithmetic mean. The calculations are shown in Table 2.5.

Table 2.5 **Translated sample for duration between successive menstrual periods in college-age women**

Value	Frequency	Value	Frequency	Value	Frequency
−4	5	1	96	6	7
−3	10	2	63	7	3
−2	28	3	24	8	2
−1	64	4	9	9	1
0	185	5	2	10	1

Note: $\bar{y} = [(-4)(5) + (-3)(10) + \cdots + (10)(1)] / 500 = 0.54$

$\bar{x} = \bar{y} + 28 = 0.54 + 28 = 28.54$ days

Similarly, systolic blood-pressure scores are usually between 100 and 200. It is easier to subtract 100 from each blood-pressure score, find the mean of the translated sample, and add 100 to obtain the mean of the original sample.

What happens to the arithmetic mean if the units or scale being worked with change? A **rescaled sample** can be created:

$$y_i = cx_i, \quad i = 1, \ldots, n$$

The following result holds:

Equation 2.2

If $y_i = cx_i, i = 1, \ldots, n$

then $\bar{y} = c\bar{x}$

Therefore, to find the arithmetic mean of the y's compute the arithmetic mean of the x's and multiply it by the constant c.

Example 2.12 Express the mean birthweight for the data in Table 2.1 in ounces rather than grams.

Solution We know that 1 oz = 28.35 g and that $\bar{x} = 3166.9$ g. Thus, if the data were expressed in terms of ounces,

$$c = \frac{1}{28.35} \quad \text{and} \quad \bar{y} = \frac{1}{28.35}(3166.9) = 111.71 \text{ oz}$$

Sometimes we want to change both the origin and the scale of the data at the same time. To do this, apply Equations 2.1 and 2.2 as follows:

Equation 2.3 Let $x_1, \ldots, x_n$ be the original sample of data and let $y_i = c_1 x_i + c_2$, $i = 1, \ldots, n$, represent a transformed sample obtained by multiplying each original sample point by a factor c_1 and then shifting over by a constant c_2.

If $y_i = c_1 x_i + c_2, \quad i = 1, \ldots, n$
then $\bar{y} = c_1 \bar{x} + c_2$

Example 2.13 If we have a sample of temperatures in °C with an arithmetic mean of 11.75°C, then what is the arithmetic mean in °F?

Solution Let y_i denote the °F temperature that corresponds to a °C temperature of x_i. The required transformation to convert the data to °F would be

$$y_i = \frac{9}{5} x_i + 32, \quad i = 1, \ldots, n$$

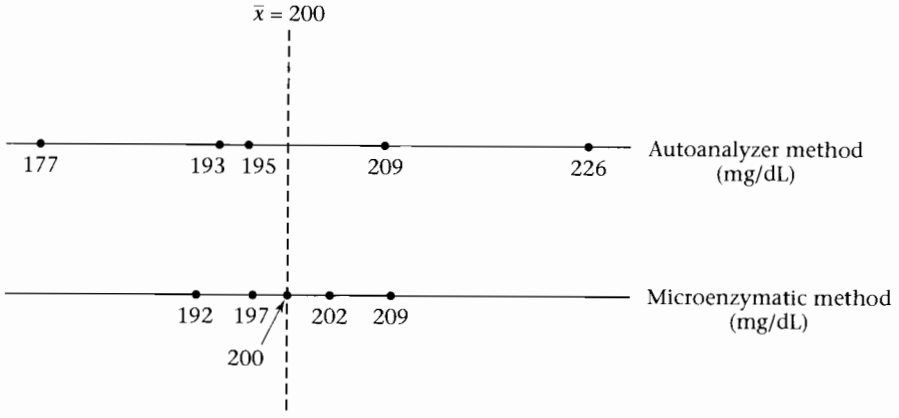
so the arithmetic mean would be

$$\bar{y} = \frac{9}{5}(11.75) + 32 = 53.15°\text{F}$$

2.4 Measures of Spread

Consider the two samples shown in Figure 2.4. They represent two samples of cholesterol measurements, each on the same person, but using different measurement techniques. They appear to have about the same center, and whatever measure of central location is used is probably about the same in the two samples. In fact, the arithmetic means are both 200 mg/dL. However, the two samples visually appear radically different. This difference lies in the greater **variability**, or **spread**, of the Autoanalyzer

Figure 2.4 **Two samples of cholesterol measurements on a given person using an Autoanalyzer and a Microenzymatic measurement technique**



method relative to the Microenzymatic method. In this section, the notion of variability is quantified. Many samples can be well described by a combination of a measure of location and a measure of spread.

The Range

Several different measures can be used to describe the variability of a sample. Perhaps the simplest measure is the range.

Definition 2.5 The **range** is the difference between the largest and smallest observations in a sample.

Example 2.14 The range in the sample of birthweights in Table 2.1 is

$$4146 - 2069 = 2077 \text{ g}$$

Example 2.15 Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4, and compare the variability of the two methods.

Solution The range for the Autoanalyzer method = 226 – 177 = 49 mg/dL. The range for the Microenzymatic method = 209 – 192 = 17 mg/dL. The Autoanalyzer method clearly seems more variable.

One advantage of the range is that it is very easy to compute once the sample points are ordered. One striking disadvantage is that it is very sensitive to extreme observations. Hence, if the lightest infant in Table 2.1 weighed 500 g rather than 2069 g, then the range would increase dramatically to 4146 – 500 = 3646 g. Another disadvantage of the range is that it depends on the sample size (n). That is, the larger n is, the larger the range tends to be. This complication makes it difficult to compare ranges from different-size data sets.

Quantiles

Another approach that addresses some of the shortcomings of the range in quantifying the spread in a data set is the use of **quantiles** or **percentiles.** Intuitively, the pth percentile is the value V_p such that p percent of the sample points are less than or equal to V_p. The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the pth percentile, depending on whether $np/100$ is an integer or not.

Definition 2.6 The ***p*th percentile** is defined by

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$)

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

Percentiles are also sometimes called **quantiles.**

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentiles have the advantage over the range of being less sensitive to outliers and of not being greatly affected by the sample size (n).

Example 2.16 Compute the 10th and 90th percentiles for the birthweight data in Table 2.1.

Solution Because $20 \times .1 = 2$ and $20 \times .9 = 18$ are integers, the 10th and 90th percentiles are defined by

10th percentile: average of the 2nd and 3rd largest values
= (2581 + 2759)/2 = 2670 g

90th percentile: average of the 18th and 19th largest values
= (3609 + 3649)/2 = 3629 g

We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution.

Example 2.17 Compute the 20th percentile for the white-blood-count data in Table 2.2.

Solution Because $np/100 = 9 \times .2 = 1.8$ is not an integer, the 20th percentile is defined by the (1 + 1)th largest value = 2nd largest value = 5000.

To compute percentiles, the sample points must be ordered. This can be difficult if n is even moderately large. An easy way to accomplish this is to use a stem-and-leaf plot (see Section 2.8), or a computer program.

There is no limit to the number of percentiles that can be computed. The most useful percentiles are often determined by the sample size and by subject-matter considerations. Frequently used percentiles are quartiles (25th, 50th, and 75th percentiles), quintiles (20th, 40th, 60th, and 80th percentiles), and deciles (10th, 20th, . . . , 90th percentiles). It is almost always instructive to look at some of the quantiles to get an overall impression of the spread and the general shape of a distribution.

The Variance and Standard Deviation

The main difference between the Autoanalyzer- and Microenzymatic-method data in Figure 2.4 is that the Microenzymatic-method values are closer to the center of the sample than the Autoanalyzer-method values. If the center of the sample is defined as the arithmetic mean, then a measure that can summarize the difference (or deviations) between the individual sample points and the arithmetic mean is needed; that is,

$$x_1 - \bar{x},\ x_2 - \bar{x},\ \ldots,\ x_n - \bar{x}$$

One simple measure that would seem to accomplish this goal is

$$d = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{n}$$

Unfortunately, this measure will not work, because of the following principle:

Equation 2.4 The sum of the deviations of the individual observations of a sample about the sample mean is always 0.

Example 2.18 Compute the sum of the deviations about the mean for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

Solution For the Autoanalyzer-method data,

$$\begin{aligned} d &= (177 - 200) + (193 - 200) + (195 - 200) + (209 - 200) + (226 - 200) \\ &= -23 - 7 - 5 + 9 + 26 = 0 \end{aligned}$$

For the Microenzymatic-method data,

$$\begin{aligned} d &= (192 - 200) + (197 - 200) + (200 - 200) + (202 - 200) + (209 - 200) \\ &= -8 - 3 + 0 + 2 + 9 = 0 \end{aligned}$$

Thus d does not help distinguish the difference in spreads between the two methods.

A second possible measure is

$$\sum_{i=1}^{n} | x_i - \bar{x} | / n$$

which is called the mean deviation. The mean deviation is a reasonable measure of spread but does not characterize the spread as well as the standard deviation (see Definition 2.8) if the underlying distribution is bell-shaped.

A third idea is to use the average of the squares of the deviations from the sample mean rather than the deviations themselves. The resulting measure of spread, denoted by s^2, is

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

The more usual form for this measure is with $n - 1$ in the denominator rather than with n. The resulting measure is called the *sample variance* (or *variance*).

Definition 2.7 The **sample variance**, or **variance**, is defined as follows:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

A rationale for using $n - 1$ in the denominator rather than n is presented in the discussion of estimation in Chapter 6.

Another commonly used measure of spread is the sample standard deviation.

Definition 2.8 The **sample standard deviation**, or **standard deviation**, is defined as follows:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\text{sample variance}}$$

Example 2.19 Compute the variance and standard deviation for the Autoanalyzer- and Microenzymatic-method data in Figure 2.4.

Solution

Autoanalyzer Method

$$\begin{aligned} s^2 &= \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4 \\ &= (529 + 49 + 25 + 81 + 676)/4 = 1360/4 = 340 \\ s &= \sqrt{340} = 18.4 \end{aligned}$$

Microenzymatic Method

$$\begin{aligned} s^2 &= \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4 \\ &= (64 + 9 + 0 + 4 + 81)/4 = 158/4 = 39.5 \\ s &= \sqrt{39.5} = 6.3 \end{aligned}$$

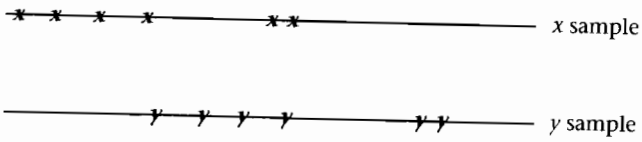
Thus the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

2.5 Some Properties of the Variance and Standard Deviation

The same question can be asked of the variance and standard deviation as of the arithmetic mean: namely, how are the variance and standard deviation affected by a change in origin or a change in the units being worked with? Suppose there is a sample $x_1, \ldots, x_n$ and all data points in the sample are shifted by a constant c; that is, a new sample $y_1, \ldots, y_n$ is created such that $y_i = x_i + c$, $i = 1, \ldots, n$.

In Figure 2.5, we would clearly expect the variance and standard deviation to remain the same, because the relationship of the points in the sample relative to one another remains the same. This property is stated as follows:

Figure 2.5 Comparison of the variances of two samples, where one sample has an origin shifted relative to the other



Equation 2.5 Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = x_i + c, \quad i = 1, \ldots, n$

If the respective sample variances of the two samples are denoted by

$$s_x^2 \quad \text{and} \quad s_y^2$$

then $s_y^2 = s_x^2$

Example 2.20 Compare the variances and standard deviations for the menstrual-period data in Tables 2.3 and 2.5.

Solution The variance and standard deviation of the two samples are the same, because the second sample was obtained from the first by subtracting 28 days from each data value; that is,

$$y_i = x_i - 28$$

Suppose the units are now changed so that a new sample $y_1, \ldots, y_n$ is created such that $y_i = cx_i$, $i = 1, \ldots, n$. The following relationship holds between the variances of the two samples.

Equation 2.6 Suppose there are two samples

$$x_1, \ldots, x_n \quad \text{and} \quad y_1, \ldots, y_n$$

where $y_i = cx_i, \quad i = 1, \ldots, n, \quad c > 0$

Then $s_y^2 = c^2 s_x^2 \quad s_y = cs_x$

This can be shown by noting that

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} (cx_i - c\bar{x})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} [c(x_i - \bar{x})]^2}{n-1} = \frac{\sum_{i=1}^{n} c^2 (x_i - \bar{x})^2}{n-1}$$

$$= \frac{c^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = c^2 s_x^2$$

$$s_y = \sqrt{c^2 s_x^2} = cs_x$$

Example 2.21 Compute the variance and standard deviation of the birthweight data in Table 2.1 in both grams and ounces.

Solution The original data are given in grams, so first compute the variance and standard deviation in these units.

$$s^2 = \frac{(3265 - 3166.9)^2 + \cdots + (2834 - 3166.9)^2}{19}$$

$$= 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$

$$s = 445.3 \text{ g}$$

To compute the variance and standard deviation in ounces, note that

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

Thus $$s^2(\text{oz}) = \frac{1}{28.35^2} s^2(\text{g}) = 246.8 \text{ oz}^2$$

$$s(\text{oz}) = \frac{1}{28.35} s(\text{g}) = 15.7 \text{ oz}$$

Thus, if the sample points change in scale by a factor of c, the variance changes by a factor of c^2 and the standard deviation changes by a factor of c. This relationship is the main reason why the standard deviation is more often used than the variance as a measure of spread, because the standard deviation and the arithmetic mean are in the same units, whereas the variance and the arithmetic mean are not. Thus, as illustrated in Examples 2.12 and 2.21, both the mean and the standard deviation change by a factor of 1/28.35 in the birthweight data of Table 2.1 when the units are expressed in ounces rather than grams.

The mean and standard deviation are the most widely used measures of location and spread in the literature. One of the main reasons for this is that the normal (or bell-shaped) distribution is defined explicitly in terms of these two parameters, and this distribution has wide applicability in many biological and medical settings. The normal distribution is discussed extensively in Chapter 5.

2.6 The Coefficient of Variation

It is useful to relate the arithmetic mean and the standard deviation to each other, because, for example, a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 than if it is 1000. A special measure, the coefficient of variation, is often used for this purpose.

Definition 2.9 The **coefficient of variation** (*CV*) is defined by

$$100\% \times (s/\bar{x})$$

This measure remains the same regardless of what units are used, because if the units change by a factor c, both the mean and standard deviation change by the factor c; the *CV*, which is the ratio between them, remains unchanged.

Example 2.22 Compute the coefficient of variation for the data in Table 2.1 when the birthweights are expressed in either grams or ounces.

Table 2.6 Reproducibility of cardiovascular risk factors in children, Bogalusa Heart Study, 1978–1979

	n	Mean	*sd*	*CV* (%)
Height (cm)	364	142.6	0.31	0.2
Weight (kg)	365	39.5	0.77	1.9
Triceps skin fold (mm)	362	15.2	0.51	3.4
Systolic blood pressure (mm Hg)	337	104.0	4.97	4.8
Diastolic blood pressure (mm Hg)	337	64.0	4.57	7.1
Total cholesterol (mg/dL)	395	160.4	3.44	2.1
HDL cholesterol (mg/dL)	349	56.9	5.89	10.4

Solution

$$CV = 100\% \times (s/\bar{x}) = 100\% \times (445.3 \text{ g}/3166.9 \text{ g}) = 14.1\%$$

If the data were expressed in ounces, then

$$CV = 100\% \times (15.7 \text{ oz}/111.71 \text{ oz}) = 14.1\%$$

The coefficient of variation is most useful in comparing the variability of several different samples, each with different arithmetic means. This is because a higher variability is usually expected when the mean increases, and the *CV* is a measure that accounts for this variability. Thus, if we are conducting a study where air pollution is measured at several sites and we wish to compare day-to-day variability at the different sites, we might expect a higher variability for the more highly polluted sites. A more accurate comparison could be made by comparing the *CV*'s at different sites than by comparing the standard deviations.

The coefficient of variation is also useful for comparing the reproducibility of different variables. Consider, for example, data from the Bogalusa Heart Study, a large study of cardiovascular risk factors in children [3], which began in the 1970s and continues up to the present time in 2004.

Children in the study were seen at approximately 3-year intervals. In 1978, a subset of the children had replicate measurements a short time apart of cardiovascular risk factors. Table 2.6 presents reproducibility data on a selected subset of cardiovascular risk factors. We note that the coefficient of variation ranges from 0.2% for height to 10.4% for HDL cholesterol. The standard deviations reported here are within-subject standard deviations based on repeated assessments on the same child. Details on how within- and between-subject variation are computed are covered in Chapter 12 when we discuss the random-effects analysis-of-variance model.

REVIEW

REVIEW QUESTIONS 2A

1. When is it appropriate to use the arithmetic mean as opposed to the median?
2. How does the geometric mean differ from the arithmetic mean? What type of data is the geometric mean used for?
3. What is the difference between the standard deviation and the coefficient of variation? When is it appropriate to use each measure?

Table 2.7 Sample of birthweights (oz) from 100 consecutive deliveries

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

2.7 Grouped Data

Sometimes the sample size is too large to display all the raw data. Also, data are frequently collected in grouped form, because the required degree of accuracy to measure a quantity exactly is often lacking, because of either measurement error or imprecise patient recall. For example, systolic blood-pressure measurements taken with a standard cuff are usually specified to the nearest 2 mm Hg, because assessing them with any more precision is difficult using this instrument. Thus, a stated measurement of 120 mm Hg may actually imply that the reading is some number ≥119 mm Hg and <121 mm Hg. Similarly, because dietary recall is generally not very accurate, the most precise estimate of fish consumption might take the following form: 2–3 servings per day, 1 serving per day, 5–6 servings per week, 2–4 servings per week, 1 serving per week, <1 serving per week and ≥1 serving per month, never.

Consider the data set in Table 2.7, which represents the birthweights from 100 consecutive deliveries at a Boston hospital. Suppose we wish to display these data for publication purposes. How can we do this? The simplest way to display the data is to generate a frequency distribution using a statistical package.

Definition 2.10 A **frequency distribution** is an ordered display of each value in a data set together with its **frequency**; that is, the number of times that value occurs in the data set. In addition, the percentage of sample points that take on a particular value is also typically given.

A frequency distribution of the sample of 100 birthweights in Table 2.7, generated using the MINITAB package, is displayed in Table 2.8.

The MINITAB Tally program provides the frequency (COUNT), relative frequency (PERCENT), cumulative frequency (CUMCNT), and cumulative percent (CUMPCT) for each birthweight present in the sample. For any particular birthweight b, the cumulative frequency, or CUMCNT, is the number of birthweights in the sample that are less than or equal to b. The PERCENT = 100 × COUNT/n, whereas CUMPCT = 100 × CUMCNT/n = the percentage of birthweights less than or equal to b.

Table 2.8 Frequency distribution of birthweight data in Table 2.7, using the MINITAB Tally program

birthwgt	COUNT	PERCENT	CUMCNT	CUMPCT
32	1	1.00	1	1.00
58	1	1.00	2	2.00
64	1	1.00	3	3.00
67	1	1.00	4	4.00
68	1	1.00	5	5.00
83	1	1.00	6	6.00
85	2	2.00	8	8.00
86	1	1.00	9	9.00
87	1	1.00	10	10.00
88	2	2.00	12	12.00
89	3	3.00	15	15.00
91	1	1.00	16	16.00
92	1	1.00	17	17.00
93	1	1.00	18	18.00
94	2	2.00	20	20.00
95	1	1.00	21	21.00
96	1	1.00	22	22.00
98	3	3.00	25	25.00
99	1	1.00	26	26.00
100	1	1.00	27	27.00
101	1	1.00	28	28.00
102	1	1.00	29	29.00
103	1	1.00	30	30.00
104	5	5.00	35	35.00
105	2	2.00	37	37.00
106	1	1.00	38	38.00
107	1	1.00	39	39.00
108	4	4.00	43	43.00
109	2	2.00	45	45.00
110	2	2.00	47	47.00
111	1	1.00	48	48.00
112	3	3.00	51	51.00
113	1	1.00	52	52.00
115	6	6.00	58	58.00
116	1	1.00	59	59.00
118	2	2.00	61	61.00
119	1	1.00	62	62.00
120	1	1.00	63	63.00
121	3	3.00	66	66.00
122	4	4.00	70	70.00
123	1	1.00	71	71.00
124	4	4.00	75	75.00
125	2	2.00	77	77.00
126	1	1.00	78	78.00
127	2	2.00	80	80.00
128	2	2.00	82	82.00
132	3	3.00	85	85.00
133	2	2.00	87	87.00
134	1	1.00	88	88.00
135	2	2.00	90	90.00
137	1	1.00	91	91.00
138	3	3.00	94	94.00
140	1	1.00	95	95.00
141	1	1.00	96	96.00
144	1	1.00	97	97.00
146	1	1.00	98	98.00
155	1	1.00	99	99.00
161	1	1.00	100	100.00
N =	100			

Table 2.9 **General layout of grouped data**

Group interval	Frequency
$y_1 \le x < y_2$	f_1
$y_2 \le x < y_3$	f_2
.	.
.	.
.	.
$y_i \le x < y_{i+1}$	f_i
.	.
.	.
.	.
$y_k \le x < y_{k+1}$	f_k

Table 2.10 **Grouped frequency distribution of birthweight (oz) from 100 consecutive deliveries**

Group interval	Frequency
$29.5 \le x < 69.5$	5
$69.5 \le x < 89.5$	10
$89.5 \le x < 99.5$	11
$99.5 \le x < 109.5$	19
$109.5 \le x < 119.5$	17
$119.5 \le x < 129.5$	20
$129.5 \le x < 139.5$	12
$139.5 \le x < 169.5$	6
	100

Note: If birthweight can only be measured to an accuracy of 0.1oz, then a possible alternate representation of the group intervals in Table 2.10 could be 29.5–69.4, 69.5–89.4, . . . , etc.

If the number of unique sample values is large, then a frequency distribution may still be too detailed a summary for publication purposes. Instead, the data could be grouped into broader categories. Here are some general instructions for categorizing the data:

(1) Subdivide the data into k intervals, starting at some lower bound y_1 and ending at some upper bound y_{k+1}.

(2) The first interval is from y_1 inclusive to y_2 exclusive; the second interval is from y_2 inclusive to y_3 exclusive; . . . ; the kth and last interval is from y_k inclusive to y_{k+1} exclusive. The rationale for this representation is to make certain the group intervals include all possible values *and* do not overlap. In some published work, grouped data are presented, but the group boundaries are ambiguous (e.g., 0–5, 5–10, etc.).

(3) The group intervals are generally chosen to be equal, although the appropriateness of equal group sizes should be dictated more by subject-matter considerations. Thus, equal intervals might be appropriate for the blood-pressure or birthweight data but not for the dietary-recall data, where the nature of the data dictates unequal group sizes corresponding to how most people remember what they eat.

(4) A count is made of the number of units that fall in each interval, which is denoted by the frequency within that interval.

(5) Finally, the group intervals and their frequencies, f_i, are then displayed concisely in a table such as Table 2.9.

For example, the raw data in Table 2.7 might be displayed in grouped form as shown in Table 2.10.

2.8 Graphic Methods

In Sections 2.1–2.7 we concentrated on methods for describing data in numeric and tabular form. In this section, these techniques are supplemented by presenting certain commonly used graphic methods for displaying data. The purpose of using graphic displays is to give a quick overall impression of the data, which is sometimes difficult to obtain with numeric measures.

Bar Graphs

One of the most widely used methods for displaying grouped data is the bar graph.

A **bar graph** can be constructed as follows:

(1) The data are divided in a number of groups using the guidelines provided in Section 2.7.

(2) For each group a rectangle is constructed with a base of a constant width and a height proportional to the frequency within that group.

(3) The rectangles are generally not contiguous and are equally spaced from each other.

A bar graph of daily vitamin-A consumption among 200 cancer cases and 200 age- and sex-matched controls is shown in Figure 2.1.

Stem-and-Leaf Plots

Two problems with bar graphs are that (1) they are somewhat difficult to construct and (2) the sense of what the actual sample points are within the respective groups is lost. One type of graphic display that overcomes these problems is the stem-and-leaf plot.

A **stem-and-leaf** plot can be constructed as follows:

(1) Separate each data point into a stem component and a leaf component, respectively, where the stem component consists of the number formed by all but the rightmost digit of the number, and the leaf component consists of the rightmost digit. Thus the stem of the number 483 is 48, and the leaf is 3.

(2) Write the smallest stem in the data set in the upper left-hand corner of the plot.

(3) Write the second stem, which equals the first stem + 1, below the first stem.

(4) Continue with step 3 until you reach the largest stem in the data set.

(5) Draw a vertical bar to the right of the column of stems.

(6) For each number in the data set, find the appropriate stem and write the leaf to the right of the vertical bar.

Figure 2.6 Stem-and-leaf plot for the birthweight data (oz) in Table 2.7

Stem-and-leaf of birthwgt N = 100
Leaf Unit = 1.0

1	3	2
1	4	
2	5	8
5	6	478
5	7	
15	8	3556788999
26	9	12344568889
45	10	0123444445567888899
(17)	11	00122235555556889
38	12	01112222344445567788
18	13	222334557888
6	14	0146
2	15	5
1	16	1

The collection of leaves thus formed takes on the general shape of the distribution of the sample points. Furthermore, the actual sample values are preserved and yet there is a grouped display for the data, which is a distinct advantage over a bar graph. Finally, it is also easy to compute the median and other quantiles from a stem-and-leaf plot. A stem-and-leaf plot is given in Figure 2.6 using MINITAB for the birthweight data in Table 2.7. Thus the point 5|8 represents 58, 11|8 represents 118, and so forth. Notice how this plot gives an overall feel for the distribution without losing the individual values. Also, the cumulative frequency count from either the lowest or the highest value is given in the first column. For the 11 stem, the absolute count is given in parentheses (17) instead of the cumulative total, because the highest or lowest value would exceed 50% (50).

In some variations of stem-and-leaf plots the leaf can consist of more than one digit. This might be appropriate for the birthweight data in Table 2.1, because the number of three-digit stems required would be very large relative to the number of data points. In this case, the leaf would consist of the rightmost two digits and the stem the leftmost two digits, and the pairs of digits to the right of the vertical bar would be underlined to distinguish between two different leaves. The stem-and-leaf display for the data in Table 2.1 is shown in Figure 2.7.

Another common variation on the ordinary stem-and-leaf plot if the number of leaves is large is to allow more than one line for each stem. Similarly, one can position the largest stem at the top of the plot and the smallest stem at the bottom of the plot. In Figure 2.8 some graphic displays using the SAS UNIVARIATE procedure illustrate this technique.

Notice that each stem is allowed two lines, with the leaves from 5 to 9 on the upper line and the leaves from 0 to 4 on the lower line. Furthermore, the leaves are

Figure 2.7 Stem-and-leaf plot for the birthweight data (g) in Table 2.1

20	69
21	
22	
23	
24	
25	81
26	
27	59
28	41 38 34
29	
30	31
31	01
32	65 60 45 00 48
33	23 14
34	84
35	41
36	49 09
37	
38	
39	
40	
41	46

ordered on each line, and a count of the number of leaves on each line is given under the # column to allow easy computation of the median and other quantiles. Thus the number 7 in the # column on the upper line for stem 12 indicates there are 7 birthweights from 125 to 129 oz in the sample, whereas the number 13 indicates there are 13 birthweights from 120 to 124 oz. Finally, a multiplication factor in the bottom of the display allows for representing decimal numbers in stem-and-leaf form. In particular, if no multiplication factor (m) is present, then it is assumed all numbers have actual value stem.leaf; whereas if m is present, then the actual value of the number is assumed to be stem.leaf $\times 10^m$. Thus, for example, because the multiplication factor is 10^1, the value 6 4 on the stem-and-leaf plot represents the number $6.4 \times 10^1 = 64$ oz.

Box Plots

In the section beginning on page 12 we discussed the comparison of the arithmetic mean and the median as a method for looking at the skewness of a distribution. This goal can also be achieved by a graphic technique known as the **box plot**. A box plot

Figure 2.8 Stem-and-leaf and box plots for the birthweight data (oz) in Table 2.7 as generated by the SAS UNIVARIATE procedure

```
Stem | Leaf           #    Boxplot
  16 | 1              1       |
  15 | 5              1       |
  15 |                        |
  14 | 6              1       |
  14 | 014            3       |
  13 | 557888         6       |
  13 | 222334         6       |
  12 | 5567788        7       |
  12 | 0111222234444  13      +------+
  11 | 5555556889     10      |      |
  11 | 0012223        7       *--+---*
  10 | 5567888899     10      |      |
  10 | 012344444      9       |      |
   9 | 568889         6       +------+
   9 | 12344          5       |
   8 | 556788999      9       |
   8 | 3              1       |
   7 |                        |
   7 |                        |
   6 | 78             2       |
   6 | 4              1       |
   5 | 8              1       0
   5 |
   4 |
   4 |
   3 |
   3 | 2              1       0
     ----+----+----+----+
Multiply Stem.Leaf by 10**+1
```

uses the relationships among the median, upper quartile, and lower quartile to describe the skewness of a distribution.

The upper and lower quartiles can be thought of conceptually as the approximate 75th and 25th percentiles of the sample—that is, the points 3/4 and 1/4 along the way in the ordered sample.

How can the median, upper quartile, and lower quartile be used to judge the symmetry of a distribution?

(1) If the distribution is symmetric, then the upper and lower quartiles should be approximately equally spaced from the median.

(2) If the upper quartile is farther from the median than the lower quartile, then the distribution is positively skewed.

(3) If the lower quartile is farther from the median than the upper quartile, then the distribution is negatively skewed.

These relationships are illustrated graphically in a box plot. In Figure 2.8 the top of the box corresponds to the upper quartile, whereas the bottom of the box corresponds to the lower quartile. A horizontal line is also drawn at the median value. Furthermore, in the SAS implementation of the box plot, the sample mean is indicated by a + sign distinct from the edges of the box.

Example 2.23 What can we learn about the symmetry properties of the distribution of birthweights from the box plot in Figure 2.8?

Solution In Figure 2.8, because the lower quartile is farther from the median than the upper quartile, the distribution is slightly negatively skewed. This pattern is true of many birthweight distributions.

In addition to displaying the symmetry properties of a sample, a box plot can also be used to visually describe the spread of a sample and can help identify possible outlying values—that is, values that seem inconsistent with the rest of the points in the sample. In the context of box plots, outlying values are defined as follows:

Definition 2.11 An **outlying value** is a value x such that either

(1) $x >$ upper quartile + 1.5 × (upper quartile − lower quartile) or

(2) $x <$ lower quartile − 1.5 × (upper quartile − lower quartile)

Definition 2.12 An **extreme outlying value** is a value x such that either

(1) $x >$ upper quartile + 3.0 × (upper quartile − lower quartile) or

(2) $x <$ lower quartile − 3.0 × (upper quartile − lower quartile)

The box plot is then completed by

(1) Drawing a vertical bar from the upper quartile to the largest nonoutlying value in the sample

(2) Drawing a vertical bar from the lower quartile to the smallest nonoutlying value in the sample

(3) Individually identifying the outlying and extreme outlying values in the sample by 0's and *'s, respectively

---

Example 2.24 Using the box plot in Figure 2.8, comment on the spread of the sample in Table 2.7 and the presence of outlying values.

Solution It can be shown from Definition 2.6 that the upper and lower quartiles are 124.5 and 98.5 oz, respectively. Hence, an outlying value x must satisfy the following relations:

$$x > 124.5 + 1.5 \times (124.5 - 98.5) = 124.5 + 39.0 = 163.5$$

or $$x < 98.5 - 1.5 \times (124.5 - 98.5) = 98.5 - 39.0 = 59.5$$

Similarly, an extreme outlying value x must satisfy the following relations:

$x > 124.5 + 3.0 \times (124.5 - 98.5) = 124.5 + 78.0 = 202.5$
or $x < 98.5 - 3.0 \times (124.5 - 98.5) = 98.5 - 78.0 = 20.5$

Thus the values 32 and 58 oz are outlying values but not extreme outlying values. These values are identified by 0's on the box plot. A vertical bar extends from 64 oz (the smallest nonoutlying value) to the lower quartile and from 161 oz (the largest nonoutlying value = the largest value in the sample) to the upper quartile. The accuracy of the two identified outlying values should probably be checked.

The methods used to identify outlying values in Definitions 2.11 and 2.12 are descriptive and unfortunately are sensitive to sample size, with more outliers detected for large sample sizes. Alternative methods for identifying outliers based on a hypothesis-testing framework are given in Chapter 8.

Many more details on stem-and-leaf plots, box plots, and other exploratory data methods are given in Tukey [4].

REVIEW QUESTIONS 2B

1 What is a stem-and-leaf plot? How does it differ from a bar graph?
2 Consider the bar graph in Figure 2.1. Is it possible to construct a stem-and-leaf plot from the data presented? If so, then construct the plot.
3 Consider the stem-and-leaf plot in Figure 2.6. Is it possible to construct a bar graph from the data presented? If so, then construct the plot.
4 What is a box plot? What additional information does this type of display give that is not available from either a bar graph or stem-and-leaf plot?

REVIEW

2.9 Case Study 1: Effects of Lead Exposure on Neurological and Psychological Function in Children

The effects of exposure to lead on the psychological and neurological well-being of children were studied [5]. Complete raw data for this study are in Data Set LEAD.DAT, and documentation for this file is in Data Set LEAD.DOC. Dr. Philip Landrigan provided this data set. All data sets are on the CD-ROM.

In summary, a group of children who lived near a lead smelter in El Paso, Texas, were identified and their blood levels of lead were measured. An exposed group of 46 children were identified who had blood-lead levels ≥ 40 μg/mL in 1972 (or for a few children in 1973). This group is defined by the variable GROUP = 2. A control group of 78 children were also identified who had blood-lead levels < 40 μg/mL in both 1972 and 1973. This group is defined by the variable GROUP = 1. All children lived close to the lead smelter.

Two important outcome variables that were studied were (1) the number of finger–wrist taps in the dominant hand (a measure of neurological function) and (2) the Wechsler full-scale IQ score. To explore the relationship of lead exposure to the

Figure 2.9 Number of finger–wrist taps in the dominant hand for exposed and control groups, El Paso Lead Study

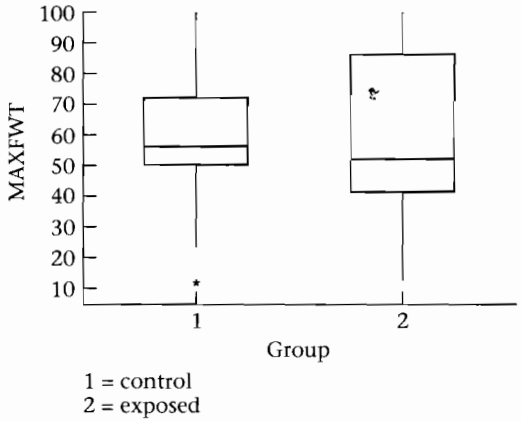
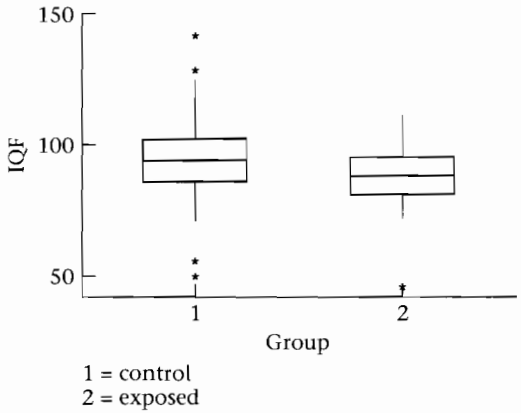


Figure 2.10 Wechsler full-scale IQ scores for exposed and control groups, El Paso Lead Study



outcome variables, we used MINITAB to obtain box plots for these two variables for children in the exposed and control groups, respectively. These are shown in Figures 2.9 and 2.10, respectively. For this purpose, because the dominant hand was not identified in the data base, we used the larger of the finger–wrist tapping scores for the right and left hand as a proxy for the number of finger–wrist taps in the dominant hand.

We note that although there is considerable spread within each group, both finger–wrist tapping scores (MAXFWT) and full-scale IQ scores (IQF) seem slightly lower in the exposed group than in the control group. We analyze these data in more detail in later chapters, using t tests, analysis of variance, and regression methods.

2.10 Case Study 2: Effects of Tobacco Use on Bone-Mineral Density in Middle-Aged Women

A twin study was performed on the relationship between bone density and cigarette consumption [6]. Forty-one pairs of middle-aged female twins who were discordant for tobacco consumption (had different smoking histories) were enrolled in a study in Australia and invited to visit a hospital in Victoria, Australia, for a measurement of bone density. Additional information was also obtained from the participants by questionnaire, including details of tobacco use; alcohol, coffee, and tea consumption; intake of calcium from dairy products; menopausal, reproductive, and fracture history; use of oral contraceptives or estrogen replacement therapy; and assessment of physical activity. Dr. John Hopper provided the data set for this study, which is available on the CD-ROM under the file name BONEDEN.DAT with documentation in BONEDEN.DOC.

Tobacco consumption was expressed in terms of *pack-years*. One pack-year is defined as 1 pack of cigarettes per day (usually about 20 cigarettes per pack) consumed

for 1 year. One advantage of using twins is that genetic influences on bone density are inherently controlled for. To analyze the data, the investigators first identified the heavier- and lighter-smoking twins in terms of pack-years. The lighter-smoking twin usually had 0 pack-years (indicating she had never smoked) or occasionally either smoked very few cigarettes per day and/or smoked for only a short time. The researchers then looked at the difference in bone-mineral density, or BMD (BMD heavier-smoking twin minus BMD lighter-smoking twin) (expressed as a percentage of the average bone density of the two twins) as a function of the difference in tobacco use (pack-years for the heavier-smoking twin minus pack-years for the lighter-smoking twin). BMD was assessed separately at three sites, the lumbar spine (lower back), the femoral neck (hip), and the femoral shaft (hip). A *scatter plot* showing the relationship between the difference in BMD versus the difference in tobacco use is given in Figure 2.11.

Note that for the lumbar spine an inverse relationship appears between the difference in BMD and the difference in tobacco use (a downward trend). Especially for twins with a large difference in tobacco use (≥30 pack-years), virtually all the differences in BMD are below 0, indicating the heavier-smoking twin had a lower BMD than the lighter-smoking twin. A similar relationship holds for the femoral neck. Results are less clear for the femoral shaft.

This is a classic example of a *matched-pair study,* which we discuss in detail beginning in Chapter 8. For such a study, the exposed (heavier-smoking twin) and control (lighter-smoking twin) are matched on other characteristics related to the outcome (BMD). In this case, the matching is based on having similar genes. We analyze this data set in more detail in later chapters, using methods based on the binomial distribution, *t* tests, and regression analysis.

2.11 Obtaining Descriptive Statistics on the Computer

Numerous statistical packages can be used to obtain descriptive statistics as well as for other statistical functions used in probability, estimation, and hypothesis testing that are covered later in this book. The CD-ROM that accompanies this book explains in detail how to use Microsoft Excel to perform these functions. Read the first chapter in the CD-ROM for details on obtaining descriptive statistics using Excel. Functions available include Average (for the arithmetic mean), Median (for the median), Stdev (for the standard deviation), Var (for the variance), GeoMean (for the geometric mean), and Percentile (for obtaining arbitrary percentiles from a sample).

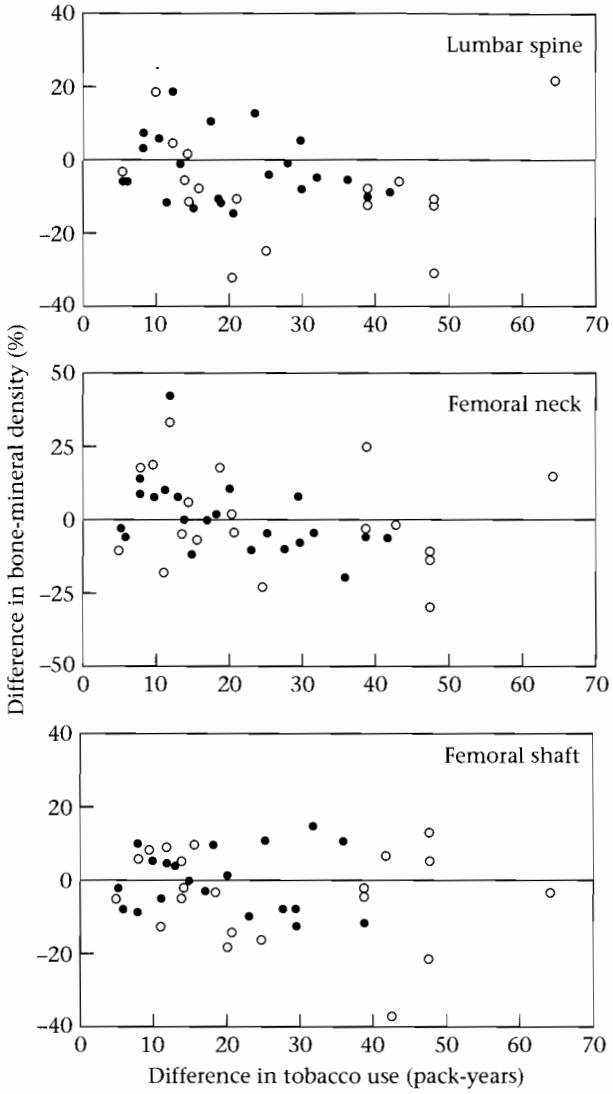
2.12 Summary

This chapter presented several **numeric and graphic methods for describing data.** These techniques are used to

(1) quickly summarize a data set

(2) and/or present results to others

Figure 2.11 **Within-pair differences in bone density at the lumbar spine, femoral neck, and femoral shaft as a function of within-pair differences in pack-years of tobacco use in 41 pairs of female twins. Monozygotic (identical) twins are represented by solid circles, and dizygotic (fraternal) twins by open circles. The difference in bone density between members of a pair is expressed as the percentage of the mean bone density for the pair.**



Source: From "The bone density of female twins discordant for tobacco use," by J. H. Hopper and E. Seeman, 1994, *The New England Journal of Medicine, 330,* 387–392. Copyright © 1994 Massachusetts Medical Society. All rights reserved.

In general, a data set can be described numerically in terms of a **measure of location** and a **measure of spread.** Several alternatives were introduced, including the **arithmetic mean, median, mode,** and **geometric mean,** as possible choices for measures of location, and the **standard deviation, quantiles,** and **range** as possible choices for measures of spread. Criteria were discussed for choosing the appropriate measures in particular circumstances. Several graphic techniques for summarizing data, including traditional methods, such as the **bar graph,** and more modern methods characteristic of exploratory data analysis (EDA), such as the **stem-and-leaf plot** and **box plot,** were introduced.

How do the descriptive methods in this chapter fit in with the methods of statistical inference discussed later in this book? Specifically, if, based on some prespecified hypotheses, some interesting trends can be found using descriptive methods, then we need some criteria to judge how "significant" these trends are. For this purpose, several commonly used **probability models** are introduced in Chapters 3 through 5 and approaches for testing the validity of these models using the methods of **statistical inference** are explored in Chapters 6 through 14.

PROBLEMS

Infectious Disease

The data in Table 2.11 are a sample from a larger data set collected on people discharged from a selected Pennsylvania hospital as part of a retrospective chart review of antibiotic usage in hospitals [7]. The data are also given in Data Set HOSPITAL.DAT with documentation in HOSPITAL.DOC on the CD-ROM. Each data set on the CD-ROM is available in six formats: ASCII, MINITAB-readable format, Excel-readable format, SAS-readable format, SPSS-readable format, and STATA-readable format.

2.1 Compute the mean and median for the duration of hospitalization for the 25 patients.

2.2 Compute the standard deviation and range for the duration of hospitalization for the 25 patients.

2.3 It is of clinical interest to know if the duration of hospitalization is affected by whether or not a patient has received antibiotics. Answer this question descriptively using either numeric or graphic methods.

Suppose the scale for a data set is changed by multiplying each observation by a positive constant.

***2.4** What is the effect on the median?

***2.5** What is the effect on the mode?

***2.6** What is the effect on the geometric mean?

***2.7** What is the effect on the range?

*Asterisk indicates that the answer to the problem is given in the Answer Section at the back of the book.

Ophthalmology

Table 2.12 comes from a paper giving the distribution of astigmatism in 1033 young men, aged 18–22, who were accepted for military service in Great Britain [8]. Assume astigmatism is rounded to the nearest 10th of a diopter and each subject in a group has the average astigmatism within that group (e.g., for the group 0.2–0.3 diopters, the actual range is from 0.15 to 0.35 diopters, and assume that each man in the group has an astigmatism of (0.15 + 0.35)/2 = 0.25 diopters).

2.8 Compute the arithmetic mean.

2.9 Compute the standard deviation.

2.10 Construct a bar graph to display the distribution of astigmatism.

2.11 Suppose we wish to construct an upper bound for astigmatism so that no more than 5% of military recruits exceed this upper bound. What is the smallest possible value for this upper bound?

Cardiovascular Disease

The data in Table 2.13 are a sample of cholesterol levels taken from 24 hospital employees who were on a standard American diet and who agreed to adopt a vegetarian diet for 1 month. Serum-cholesterol measurements were made before adopting the diet and 1 month after.

***2.12** Compute the mean change in cholesterol.

Table 2.11 Hospital-stay data

ID no.	Duration of hospital stay	Age	Sex 1 = M 2 = F	First temp. following admission	First WBC ($\times 10^3$) following admission	Received anti-biotic 1 = yes 2 = no	Received bacterial culture 1 = yes 2 = no	Service 1 = med. 2 = surg.
1	5	30	2	99.0	8	2	2	1
2	10	73	2	98.0	5	2	1	1
3	6	40	2	99.0	12	2	2	2
4	11	47	2	98.2	4	2	2	2
5	5	25	2	98.5	11	2	2	2
6	14	82	1	96.8	6	1	2	2
7	30	60	1	99.5	8	1	1	1
8	11	56	2	98.6	7	2	2	1
9	17	43	2	98.0	7	2	2	1
10	3	50	1	98.0	12	2	1	2
11	9	59	2	97.6	7	2	1	1
12	3	4	1	97.8	3	2	2	2
13	8	22	2	99.5	11	1	2	2
14	8	33	2	98.4	14	1	1	2
15	5	20	2	98.4	11	2	1	2
16	5	32	1	99.0	9	2	2	2
17	7	36	1	99.2	6	1	2	2
18	4	69	1	98.0	6	2	2	2
19	3	47	1	97.0	5	1	2	1
20	7	22	1	98.2	6	2	2	2
21	9	11	1	98.2	10	2	2	2
22	11	19	1	98.6	14	1	2	2
23	11	67	2	97.6	4	2	2	1
24	9	43	2	98.6	5	2	2	2
25	4	41	2	98.0	5	2	2	1

Table 2.12 Distribution of astigmatism in 1033 young men aged 18–22

Degree of astigmatism (diopters)	Frequency
0.0 or less than 0.2	458
0.2–0.3	268
0.4–0.5	151
0.6–1.0	79
1.1–2.0	44
2.1–3.0	19
3.1–4.0	9
4.1–5.0	3
5.1–6.0	2
	1033

Source: Reprinted with permission of the Editor, the authors and the Journal from the *British Medical Journal*, May 7, 1394–1398, 1960.

*__2.13__ Compute the standard deviation of the change in cholesterol levels.

__2.14__ Construct a stem-and-leaf plot of the cholesterol changes.

*__2.15__ Compute the median change in cholesterol.

__2.16__ Construct a box plot of the cholesterol changes to the right of the stem-and-leaf plot.

__2.17__ Comment on the symmetry of the distribution of change scores based on your answers to Problems 2.12 through 2.16.

__2.18__ Some investigators feel the effects of diet on cholesterol are more evident in people with high rather than low cholesterol levels. If you split the data in Table 2.13 according to whether baseline cholesterol is above or below the median, can you comment descriptively on this issue?

Table 2.13 Serum-cholesterol levels (mg/dL) before and after adopting a vegetarian diet

Subject	Before	After	Difference*
1	195	146	49
2	145	155	−10
3	205	178	27
4	159	146	13
5	244	208	36
6	166	147	19
7	250	202	48
8	236	215	21
9	192	184	8
10	224	208	16
11	238	206	32
12	197	169	28
13	169	182	−13
14	158	127	31
15	151	149	2
16	197	178	19
17	180	161	19
18	222	187	35
19	168	176	−8
20	168	145	23
21	167	154	13
22	161	153	8
23	178	137	41
24	137	125	12

* Before − after.

Hypertension

An experiment looked at the effect of position on level of blood pressure [9]. In the experiment 32 participants had their blood pressures measured while lying down with their arms at their sides and again standing with their arms supported at heart level. The data are given in Table 2.14.

2.19 Compute the arithmetic mean and median for the difference in systolic and diastolic blood pressure, respectively, taken in different positions (recumbent minus standing).

2.20 Construct stem-and-leaf and box plots for the difference scores for each type of blood pressure.

2.21 Based on your answers to Problems 2.19 and 2.20, comment on the effect of position on the levels of systolic and diastolic blood pressure.

Table 2.14 Effect of position on blood pressure

	Blood pressure (mm Hg)			
Participant	Recumbent, arm at side		Standing, arm at heart level	
B. R. A.	99[a]	71[b]	105[a]	79[b]
J. A. B.	126	74	124	76
F. L. B.	108	72	102	68
V. P. B.	122	68	114	72
M. F. B.	104	64	96	62
E. H. B.	108	60	96	56
G. C.	116	70	106	70
M. M. C.	106	74	106	76
T. J. F.	118	82	120	90
R. R. F.	92	58	88	60
C. R. F.	110	78	102	80
E. W. G.	138	80	124	76
T. F. H.	120	70	118	84
E. J. H.	142	88	136	90
H. B. H.	118	58	92	58
R. T. K.	134	76	126	68
W. E. L.	118	72	108	68
R. L. L.	126	78	114	76
H. S. M.	108	78	94	70
V. J. M.	136	86	144	88
R. H. P.	110	78	100	64
R. C. R.	120	74	106	70
J. A. R.	108	74	94	74
A. K. R.	132	92	128	88
T. H. S.	102	68	96	64
O. E. S.	118	70	102	68
R. E. S.	116	76	88	60
E. C. T.	118	80	100	84
J. H. T.	110	74	96	70
F. P. V.	122	72	118	78
P. F. W.	106	62	94	56
W. J. W.	146	90	138	94

[a]Systolic blood pressure
[b]Diastolic blood pressure
Source: C. E. Kossman (1946), "Relative importance of certain variables in the clinical determination of blood pressure," *American Journal of Medicine, 1,* 464–467. Reprinted with permission of the *American Journal of Medicine.*

2.22 Orthostatic hypertension is sometimes defined based on an unusual change in blood pressure after changing position. Suppose we define a normal range for change in systolic blood pressure (SBP) based on change in SBP from the recumbent to the standing position in Table 2.14 that is between the upper and lower decile. What should the normal range be?

Pulmonary Disease

FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. Data Set FEV.DAT (on the CD-ROM) contains determinations of FEV in 1980 on 654 children ages 3–19 who were seen in the Childhood Respiratory Disease Study (CRD Study) in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children [10].

The data in Table 2.15 are available for each child.

Table 2.15 Format for FEV.DAT

Column	Variable	Format or code
1–5	ID number	
7–8	Age (years)	
10–15	FEV (liters)	X.XXX
17–20	Height (inches)	XX.X
22	Sex	0 = female/1 = male
24	Smoking status	0 = noncurrent smoker/ 1 = current smoker

2.23 For each variable (other than ID), obtain appropriate descriptive statistics (both numeric and graphic).

2.24 Use both numeric and graphic measures to assess the relationship of FEV to age, height, and smoking status. (Do this separately for boys and girls.)

2.25 Compare the pattern of growth of FEV by age for boys and girls. Are there any similarities? Any differences?

Nutrition

The food-frequency questionnaire (FFQ) is an instrument often used in dietary epidemiology to assess consumption of specific foods. A person is asked to write down number of servings per day typically eaten in the past year of over 100 individual food items. A food-composition table is then used to compute nutrient intakes (protein, fat, etc.), based on aggregating responses for individual foods. The FFQ is inexpensive to administer but is considered less accurate than the diet record (DR) (the gold standard of dietary epidemiology). For the diet record, a participant writes down the amount of each specific food eaten over the past week in a food diary and a nutritionist using a special computer program computes nutrient intakes from the food diaries. This is a much more expensive method of dietary recording. To validate the FFQ, 173 nurses participating in the Nurses' Health Study completed 4 weeks of diet recording about equally spaced over a 12-month period and an FFQ at the end of diet recording [11]. Data are presented in Data Set VALID.DAT (on the CD-ROM) for saturated fat, total fat, total alcohol consumption, and total caloric intake for both the DR and FFQ. For the DR, average nutrient intakes were computed over the 4 weeks of diet recording. Table 2.16 shows the format of this file.

Table 2.16 Format for VALID.DAT

Column	Variable	Format or code
1–6	ID number	XXXXX.XX
8–15	Saturated fat—DR (g)	XXXXX.XX
17–24	Saturated fat—FFQ (g)	XXXXX.XX
26–33	Total fat—DR (g)	XXXXX.XX
35–42	Total fat—FFQ (g)	XXXXX.XX
44–51	Alcohol consumption—DR (oz)	XXXXX.XX
53–60	Alcohol consumption—FFQ (oz)	XXXXX.XX
62–70	Total calories—DR	XXXXXX.XX
72–80	Total calories—FFQ	XXXXXX.XX

2.26 Compute appropriate descriptive statistics for each nutrient for both DR and FFQ, using both numeric and graphic measures.

2.27 Use descriptive statistics to relate nutrient intake for the DR and FFQ. Do you think the FFQ is a reasonably accurate approximation to the DR? Why or why not?

2.28 A frequently used method for quantifying dietary intake is in the form of quintiles. Compute quintiles for each nutrient and each method of recording, and relate the nutrient composition for DR and FFQ using the quintile scale. (That is, how does the quintile category based on DR relate to the quintile category based on FFQ for the same individual?) Do you get the same impression about the concordance between DR and FFQ using quintiles as in Problem 2.27, where raw (ungrouped) nutrient intake is considered?

In nutritional epidemiology, it is customary to assess nutrient intake in relationship to total caloric intake. One measure used to accomplish this is *nutrient density,* which is defined as 100% × (caloric intake of a nutrient/total caloric intake). For fat consumption, 1 gram of fat is equivalent to 9 calories.

2.29 Compute the nutrient density for total fat for the DR and FFQ, and obtain appropriate descriptive statistics for this variable. How do they compare?

2.30 Relate the nutrient density for total fat for the DR versus the FFQ using the quintile approach in Problem 2.28. Is the concordance between total fat for DR and FFQ stronger, weaker, or the same when total fat is expressed in terms of nutrient density as opposed to raw nutrient?

Environmental Health, Pediatrics

In Section 2.9, we described Data Set LEAD.DAT (on the CD-ROM) concerning the effect of lead exposure on neurological and psychological function in children.

2.31 Compare the exposed and control groups on age and gender, using appropriate numeric and graphic descriptive measures.

2.32 Compare the exposed and control groups on verbal and performance IQ, using appropriate numeric and graphic descriptive measures.

Cardiovascular Disease

Activated-protein-C (APC) resistance is a serum marker that has been associated with thrombosis (the formation of blood clots often leading to heart attacks) among adults. A study assesses this risk factor among adolescents. To assess the reproducibility of the assay, a split-sample technique was used where a blood sample was provided by 10 people; each sample was split into two aliquots (subsamples), and each aliquot was assessed separately. Table 2.17 gives the results.

2.33 Suppose the variation between split samples is thought to be a function of the mean level, where more variation is expected as the mean level increases. What measure of reproducibility can be used under these circumstances?

2.34 Compute the measure in Problem 2.33 for each participant. Then obtain an average of this measure over the 10 participants. If this average is less than 10%, it indicates excellent reproducibility. Is this true for APC resistance? *Hint:* Consider using Excel to answer this question.

Microbiology

The purpose of this study was to demonstrate that soy beans inoculated with nitrogen-fixing bacteria yield more and grow adequately without expensive environmentally deleterious synthesized fertilizers. The trial was conducted under controlled conditions with uniform amounts of soil. The initial hypothesis was that inoculated plants would outperform their uninoculated counterparts. This assumption is based on the facts that plants need nitrogen to manufacture vital proteins and amino acids and that nitrogen-fixing bacteria would make more of this substance available to plants, increasing their size and yield. There were 8 inoculated plants (I) and 8 uninoculated plants (U). The plant yield as measured by pod weight for each plant is given in Table 2.18.

Table 2.17 APC resistance split-samples data

Sample number	A	B	$A - B$
1	2.22	1.88	0.34
2	3.42	3.59	−0.17
3	3.68	3.01	0.67
4	2.64	2.37	0.27
5	2.68	2.26	0.42
6	3.29	3.04	0.25
7	3.85	3.57	0.28
8	2.24	2.29	−0.05
9	3.25	3.39	−0.14
10	3.30	3.16	0.14

2.35 Compute appropriate descriptive statistics for I and U plants.

2.36 Use graphic methods to compare the two groups.

2.37 What is your overall impression concerning the pod weight in the two groups?

Table 2.18 Pod weight (g) from inoculated (I) and uninoculated (U) plants

Sample number	I	U
1	1.76	0.49
2	1.45	0.85
3	1.03	1.00
4	1.53	1.54
5	2.34	1.01
6	1.96	0.75
7	1.79	2.11
8	1.21	0.92

Note: The data for this problem were supplied by David Rosner.

Endocrinology

In Section 2.10, we described Data Set BONEDEN.DAT (on the CD-ROM) concerning the effect of tobacco use on bone-mineral density (BMD).

2.38 For each pair of twins, for the lumbar spine, compute

A = BMD for the heavier-smoking twin − BMD for the lighter-smoking twin = $x_1 - x_2$
B = mean BMD for the twinship = $(x_1 + x_2)/2$
$C = 100\% \times (A/B)$

Derive appropriate descriptive statistics for C over the entire study population.

2.39 Suppose we group the twin pairs according to the difference in tobacco use expressed in 10 pack-year groups (0–9.9 pack-years/10–19.9 pack-years/20–29.9 pack-years/30–39.9 pack-years/40+ pack-years). Compute appropriate descriptive statistics, and provide a scatter plot for C grouped by the difference in tobacco use in pack-years.

2.40 What impression do you have of the relationship between BMD and tobacco use based on Problem 2.39?

2.41–2.43 Answer Problems 2.38–2.40 for BMD for the femoral neck.

2.44–2.46 Answer Problems 2.38–2.40 for BMD for the femoral shaft.

REFERENCES

[1] White, J. R., & Froeb, H. E. (1980). Small-airways dysfunction in nonsmokers chronically exposed to tobacco smoke. *New England Journal of Medicine, 302*(33), 720–723.

[2] Pedersen, A., Wiesner, P., Holmes, K., Johnson, C., & Turck, M. (1972). Spectinomycin and Penicillin G in the treatment of gonorrhea. *JAMA, 220*(2), 205–208.

[3] Foster, T. A., & Berenson, G. (1987). Measurement error and reliability in four pediatric cross-sectional surveys of cardiovascular disease risk factor variables—the Bogalusa Heart Study. *Journal of Chronic Disease, 40*(1), 13–21.

[4] Tukey, J. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

[5] Landrigan, P. J., Whitworth, R. H., Baloh, R. W., Staehling, N. W., Barthel, W. F., & Rosenblum, B. F. (1975, March 29). Neuropsychological dysfunction in children with chronic low-level lead absorption. *Lancet, 1,* 708–715.

[6] Hopper, J. H., & Seeman, E. (1994). The bone density of female twins discordant for tobacco use. *New England Journal of Medicine, 330,* 387–392.

[7] Townsend, T. R., Shapiro, M., Rosner, B., & Kass, E. H. (1979). Use of antimicrobial drugs in general hospitals I. Description of population and definition of methods. *Journal of Infectious Diseases, 139*(6), 688–697.

[8] Sorsby, A., Sheridan, M., Leary, G. A., & Benjamin, B. (1960, May 7). Vision, visual acuity and ocular refraction of young men in a sample of 1033 subjects. *British Medical Journal,* 1394–1398.

[9] Kossmann, C. E. (1946). Relative importance of certain variables in clinical determination of blood pressure. *American Journal of Medicine, 1,* 464–467.

[10] Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology, 110,* 15–26.

[11] Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H., & Speizer, F. E. (1985). Reproducibility and validity of a semi-quantitative food frequency questionnaire. *American Journal of Epidemiology, 122,* 51–65.

3 Probability

3.1 Introduction

Chapter 2 outlined various techniques for concisely describing data. But we usually want to do more with data than just describe them. In particular, we might want to test certain specific inferences about the behavior of the data.

Example 3.1 **Cancer** One theory concerning the etiology of breast cancer states that women in a given age group who give birth to their first child relatively late in life (after 30) are at greater risk for eventually developing breast cancer over some time period t than are women who give birth to their first child early in life (before 20). Because women in the upper social classes tend to have children later, this theory has been used to explain why these women have a higher risk of developing breast cancer than women in the lower social classes. To test this hypothesis, we might identify 2000 women from a particular census tract who are currently ages 45–54 and have never had breast cancer, of whom 1000 had their first child before the age of 20 (call this group A) and 1000 after the age of 30 (group B). These 2000 women might be followed for 5 years to assess if they developed breast cancer during this period. Suppose there are 4 new cases of breast cancer in group A and 5 new cases in group B.

Is this evidence enough to confirm a difference in risk between the two groups? Most people would feel uneasy about concluding that on the basis of such a limited amount of data.

Suppose we had a more ambitious plan and sampled 10,000 women from groups A and B, respectively, and at follow-up found 40 new cases in group A and 50 new cases in group B and asked the same question. Although we might be more comfortable with the conclusion because of the larger sample size, we would still have to admit this apparent difference in the rates could be due to chance.

The problem is that we need a conceptual framework to make these decisions but have not explicitly stated what the framework is. This framework is provided by the underlying concept of **probability**. In this chapter probability is defined and some rules for working with probabilities are introduced. Understanding probability

is essential in calculating and interpreting p-values in the statistical tests of subsequent chapters. It also permits the discussion of sensitivity, specificity, and predictive values of screening tests in Section 3.7.

3.2 Definition of Probability

Example 3.2 **Obstetrics** Suppose we are interested in the probability of a male live childbirth (or livebirth) among all livebirths in the United States. Conventional wisdom tells us this probability should be close to .5. We can explore this subject by looking at some vital-statistics data, as presented in Table 3.1 [1]. The probability of a male livebirth based on 1965 data is .51247; based on 1965–1969 data, .51248; and based on 1965–1974 data, .51268. These **empirical** probabilities are based on a finite amount of data. In principle, the sample size could be expanded indefinitely and an increasingly more precise estimate of the probability obtained.

Table 3.1 **Probability of a male livebirth during the period 1965–1974**

Time period	Number of male livebirths (a)	Total number of livebirths (b)	Empirical probability of a male livebirth (a/b)
1965	1,927,054	3,760,358	.51247
1965–1969	9,219,202	17,989,361	.51248
1965–1974	17,857,857	34,832,051	.51268

This principle leads to the following definition of probability:

Definition 3.1 The **sample space** is the set of all possible outcomes. In referring to probabilities of events, an **event** is any set of outcomes of interest. The **probability** of an event is the relative frequency (see p. 25) of this set of outcomes over an indefinitely large (or infinite) number of trials.

Example 3.3 **Pulmonary Disease** The tuberculin skin test is a routine screening test used to detect tuberculosis. The results of this test can be categorized as either positive, negative, or uncertain. If the probability of a positive test is .1, it means that if a large number of such tests were performed, about 10% would be positive. The actual percentage of positive tests will be increasingly close to .1 the more tests are performed.

Example 3.4 **Cancer** The probability of developing a new case of breast cancer in 30 years in 40-year-old women who have never had breast cancer is approximately 1/11. This probability means that over a large sample of 40-year-old women who have never had breast cancer, approximately 1 in 11 will develop the disease by age 70, with

this proportion becoming increasingly close to 1 in 11 as the number of women sampled increases.

In real life, experiments cannot be performed an infinite number of times. Instead, probabilities of events are estimated from the empirical probabilities obtained from large samples (as in Examples 3.2–3.4). In other instances, theoretical-probability models are constructed from which probabilities of many different kinds of events can be computed. An important issue in statistical inference is to compare empirical probabilities with theoretical probabilities—that is, to assess the goodness-of-fit of probability models. This topic is covered in Section 10.7.

Example 3.5 **Cancer** The probability of developing a new case of stomach cancer over a 1-year period for 45- to 49-year-old women, based on Connecticut Tumor Registry data from 1963 to 1965, is 14 per 100,000 [2]. Suppose we have studied cancer rates in a small group of Connecticut nurses over this period and want to compare how close the rates from this limited sample are to the tumor-registry figures. The value 14 per 100,000 would be the best estimate of the probability before collecting any data, and we would then see how closely our new sample data conformed with this probability.

From Definition 3.1 and from the preceding examples, we can deduce that probabilities have the following basic properties:

Equation 3.1

(1) The probability of an event E, denoted by $Pr(E)$, always satisfies $0 \le Pr(E) \le 1$.

(2) If outcomes A and B are two events that cannot both happen at the same time, then $Pr(A \text{ or } B \text{ occurs}) = Pr(A) + Pr(B)$.

Example 3.6 **Hypertension** Let A be the event that a person has normotensive diastolic blood-pressure (DBP) readings (DBP < 90), and let B be the event that a person has borderline DBP readings ($90 \le$ DBP < 95). Suppose that $Pr(A) = .7$, $Pr(B) = .1$. Let z be the event that a person has DBP < 95. Then

$$Pr(z) = Pr(A) + Pr(B) = .8$$

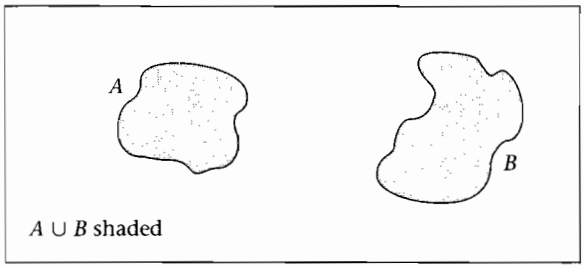
because the events A and B cannot occur at the same time.

Definition 3.2 Two events A and B are **mutually exclusive** if they cannot both happen at the same time.

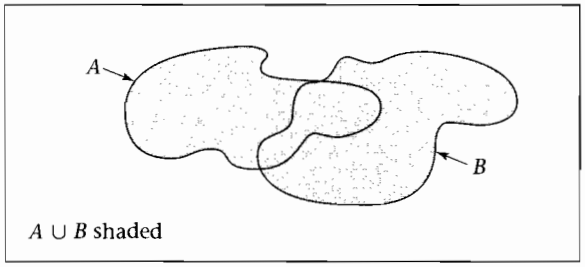
Thus the events A and B in Example 3.6 are mutually exclusive.

Example 3.7 **Hypertension** Let X be DBP, C be the event $X \ge 90$, and D be the event $75 \le X \le 100$. Events C and D are *not* mutually exclusive, because they both occur when $90 \le X \le 100$.

Figure 3.1 Diagrammatic representation of $A \cup B$: (a) A, B mutually exclusive; (b) A, B not mutually exclusive



(a)



(b)

3.3 Some Useful Probabilistic Notation

Definition 3.3 The symbol { } is used as shorthand for the phrase "the event."

Definition 3.4 $A \cup B$ is the event that either A or B occurs, or they both occur.

Figure 3.1 diagrammatically depicts $A \cup B$ both for the case where A and B are and are not mutually exclusive.

Example 3.8 **Hypertension** Let events A and B be defined as in Example 3.6: $A = \{X < 90\}$, $B = \{90 \le X < 95\}$, where X = DBP. Then $A \cup B = \{X < 95\}$.

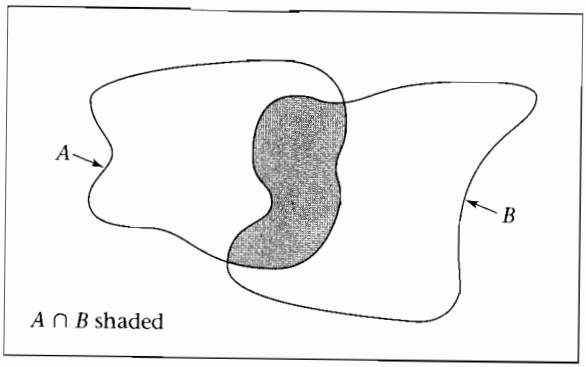
Example 3.9 **Hypertension** Let events C and D be defined as in Example 3.7:

$$C = \{X \ge 90\} \qquad D = \{75 \le X \le 100\}$$

Then $$C \cup D = \{X \ge 75\}$$

Definition 3.5 $A \cap B$ is the event that both A and B occur simultaneously. $A \cap B$ is depicted diagrammatically in Figure 3.2.

Figure 3.2 Diagrammatic representation of $A \cap B$



Example 3.10 **Hypertension** Let events C and D be defined as in Example 3.7; that is,

$$C = \{X \geq 90\} \qquad D = \{75 \leq X \leq 100\}$$

Then $$C \cap D = \{90 \leq X \leq 100\}$$

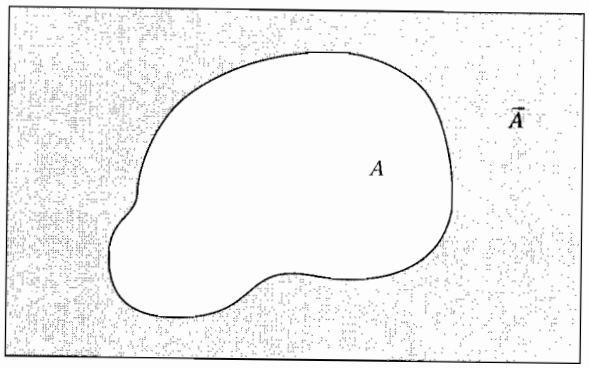
Notice that $A \cap B$ is not well defined for events A and B in Example 3.6, because both A and B cannot occur simultaneously. This situation is true for any mutually exclusive events.

Definition 3.6 $\overline{A}$ is the event that A does not occur. It is called the **complement** of A. Notice that $Pr(\overline{A}) = 1 - Pr(A)$, because $\overline{A}$ occurs only when A does not occur. Event $\overline{A}$ is diagrammed in Figure 3.3.

Example 3.11 **Hypertension** Let events A and C be defined as in Examples 3.6 and 3.7; that is,

$$A = \{X < 90\} \qquad C = \{X \geq 90\}$$

Figure 3.3 Diagrammatic representation of $\overline{A}$



Then $C = \overline{A}$, because C can only occur when A does not occur. Notice that

$$Pr(C) = Pr(\overline{A}) = 1 - .7 = .3$$

Thus, if 70% of people have diastolic blood pressure (DBP) < 90, then 30% of people must have DBP ≥ 90.

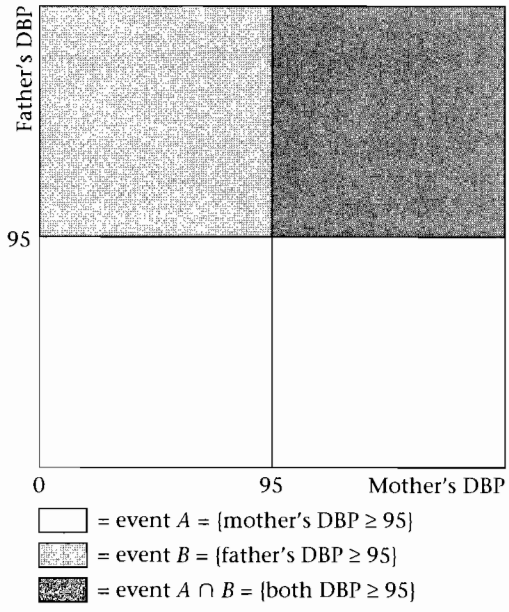
3.4 The Multiplication Law of Probability

In the preceding section, events in general were described. In this section, certain specific types of events are discussed.

Example 3.12 **Hypertension, Genetics** Suppose we are conducting a hypertension-screening program in the home. Consider all possible pairs of DBP measurements of the mother and father within a given family, assuming that the mother and father are not genetically related. This sample space consists of all pairs of numbers of the form (X, Y) where $X > 0$, $Y > 0$. Certain specific events might be of interest in this context. In particular, we might be interested in whether the mother or father is hypertensive, which is described, respectively, by events A = {mother's DBP ≥ 95}, B = {father's DBP ≥ 95}. These events are diagrammed in Figure 3.4.

Suppose we know that $Pr(A) = .1$, $Pr(B) = .2$. What can we say about $Pr(A \cap B)$ = Pr(mother's DBP ≥ 95 and father's DBP ≥ 95) = Pr(both mother and father are hypertensive)? We can say nothing unless we are willing to make certain assumptions.

Figure 3.4 **Possible diastolic blood-pressure measurements of the mother and father within a given family**



Definition 3.7 Two events A and B are called **independent events** if

$$Pr(A \cap B) = Pr(A) \times Pr(B)$$

Example 3.13 **Hypertension, Genetics** Compute the probability that both mother and father are hypertensive if the events in Example 3.12 are independent.

Solution If A and B are independent events, then

$$Pr(A \cap B) = Pr(A) \times Pr(B) = .1(.2) = .02$$

One way to interpret this example is to assume that the hypertensive status of the mother does not depend at all on the hypertensive status of the father. Thus, if these events are independent, then in 10% of all households where the father is hypertensive the mother is also hypertensive, and in 10% of all households where the father is *not* hypertensive the mother is hypertensive. We would expect these two events to be independent if the primary determinants of elevated blood pressure were genetic. However, if the primary determinants of elevated blood pressure were, to some extent, environmental, then we would expect the mother would be more likely to have elevated blood pressure (A true) if the father had elevated blood pressure (B true) than if the father did not have elevated blood pressure (B not true). In this latter case the events would not be independent. The implications of this lack of independence are discussed later in this chapter.

If two events are not independent, then they are said to be dependent.

Definition 3.8 Two events A, B are **dependent** if

$$Pr(A \cap B) \neq Pr(A) \times Pr(B)$$

Example 3.14 is a classic example of dependent events.

Example 3.14 **Hypertension, Genetics** Consider all possible diastolic blood-pressure measurements from a mother and her first-born child. Let

A = {mother's DBP ≥ 95} B = {first-born child's DBP ≥ 80}

Suppose $Pr(A \cap B) = .05$ $Pr(A) = .1$ $Pr(B) = .2$

Then $Pr(A \cap B) = .05 > Pr(A) \times Pr(B) = .02$

and the events A, B would be dependent.

This outcome would be expected, because the mother and first-born child both share the same environment and are genetically related. In other words, the first-born child is more likely to have elevated blood pressure in households where the mother is hypertensive than in households where the mother is not hypertensive.

Example 3.15 **Sexually Transmitted Disease** Suppose two doctors, A and B, test all patients coming into a VD clinic for syphilis. Let events A^+ = {doctor A makes a positive diagnosis}, B^+ = {doctor B makes a positive diagnosis}. Suppose doctor A diagnoses 10% of all patients as positive, doctor B diagnoses 17% of all patients as positive, and both doctors diagnose 8% of all patients as positive. Are the events A^+, B^+ independent?

Solution We are given that

$$Pr(A^+) = .1 \quad Pr(B^+) = .17 \quad Pr(A^+ \cap B^+) = .08$$

Thus $Pr(A^+ \cap B^+) = .08 > Pr(A^+) \times Pr(B^+) = .1(.17) = .017$

and the events are dependent. This result would be expected, because there should be a similarity between how two doctors diagnose patients for syphilis.

Definition 3.7 can be generalized to the case of $k(>2)$ independent events. This is often called the *multiplication law of probability.*

Equation 3.2 **Multiplication Law of Probability**

If $A_1, \ldots, A_k$ are mutually independent events,

then $Pr(A_1 \cap A_2 \cap \cdots \cap A_k) = Pr(A_1) \times Pr(A_2) \times \cdots \times Pr(A_k)$

3.5 The Addition Law of Probability

We have seen from the definition of probability that if A and B are mutually exclusive events, then $Pr(A \cup B) = Pr(A) + Pr(B)$. A more general formula for $Pr(A \cup B)$ can be developed when events A and B are not necessarily mutually exclusive. This formula, the *addition law of probability,* is stated as follows:

Equation 3.3 **Addition Law of Probability**

If A and B are any events,

then $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

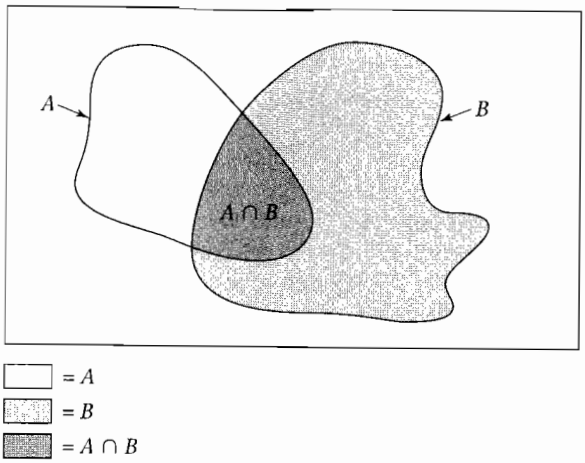
This principle is diagrammed in Figure 3.5. Thus, to compute $Pr(A \cup B)$, add the probabilities of A and B separately and then subtract the overlap, which is $Pr(A \cap B)$.

Example 3.16 **Sexually Transmitted Disease** Consider the data in Example 3.15. Suppose a patient is referred for further lab tests if either doctor A or B makes a positive diagnosis. What is the probability that a patient will be referred for further lab tests?

Solution The event that either doctor makes a positive diagnosis can be represented by $A^+ \cup B^+$. We know that

$$Pr(A^+) = .1 \quad Pr(B^+) = .17 \quad Pr(A^+ \cap B^+) = .08$$

Figure 3.5 **Diagrammatic representation of the addition law of probability**



Therefore, from the addition law of probability,

$$Pr(A^+ \cup B^+) = Pr(A^+) + Pr(B^+) - Pr(A^+ \cap B^+) = .1 + .17 - .08 = .19$$

Thus 19% of all patients will be referred for further lab tests.

Special cases of the addition law are of interest. First, if events A and B are *mutually exclusive,* then $Pr(A \cap B) = 0$ and the addition law reduces to $Pr(A \cup B) = Pr(A) + Pr(B)$. This property is given in Equation 3.1 for probabilities over any two mutually exclusive events. Second, if events A and B are *independent,* then by definition $Pr(A \cap B) = Pr(A) \times Pr(B)$ and $Pr(A \cup B)$ can be rewritten as $Pr(A) + Pr(B) - Pr(A) \times Pr(B)$. This leads to the following important special case of the addition law.

Equation 3.4 **Addition Law of Probability for Independent Events**

If two events A and B are independent, then

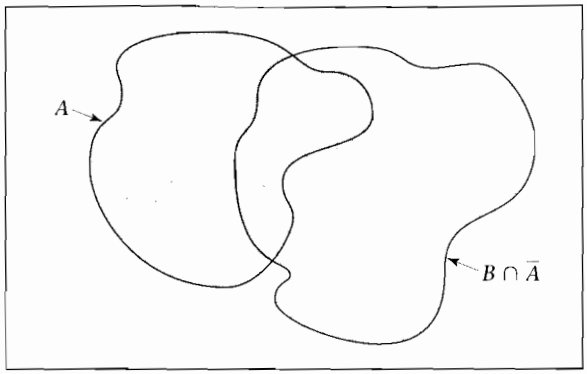
$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)]$$

This special case of the addition law can be interpreted as follows: The event $A \cup B$ can be separated into two mutually exclusive events: {A occurs} and {B occurs and A does not occur}. Furthermore, because of the independence of A and B, the probability of the latter event can be written as $Pr(B) \times [1 - Pr(A)]$. This probability is diagrammed in Figure 3.6.

Example 3.17 **Hypertension** Look at Example 3.12, where

A = {mother's DBP ≥ 95} and B = {father's DBP ≥ 95}

Figure 3.6 **Diagrammatic representation of the addition law of probability for independent events**



= A

= {B occurs and A does not occur} = $B \cap \bar{A}$

$Pr(A) = .1$, $Pr(B) = .2$, and assume A and B are independent events. Suppose a "hypertensive household" is defined as one in which either the mother or the father is hypertensive, and hypertension is defined for the mother and father, respectively, in terms of events A and B. What is the probability of a hypertensive household?

Solution Pr(hypertensive household) is

$$Pr(A \cup B) = Pr(A) + Pr(B) \times \left[1 - Pr(A)\right] = .1 + .2(.9) = .28$$

Thus 28% of all households will be hypertensive.

It is possible to extend the addition law to more than two events. In particular, if there are three events A, B, and C, then

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C)$$

This result can be generalized to an arbitrary number of events, although that is beyond the scope of this text (see [3]).

3.6 Conditional Probability

Suppose we want to compute the probability of several events occurring simultaneously. If the events are independent, then we can use the multiplication law of probability to do so. If some of the events are dependent, then a quantitative measure of dependence is needed to extend the multiplication law to the case of dependent events. Consider the following example:

Example 3.18 **Cancer** Physicians recommend that all women over age 50 be screened for breast cancer. The definitive test for identifying breast tumors is a breast biopsy. However,

this procedure is too expensive and invasive to recommend for *all* women over the age of 50. Instead, women in this age group are encouraged to have a mammogram every 1 to 2 years. Women with positive mammograms are then tested further with a biopsy. The ideal situation would be if the probability of breast cancer among women who are mammogram positive is 1, whereas the probability of breast cancer among women who are mammogram negative is 0. The two events {mammogram positive} and {breast cancer} would then be completely dependent; the results of the screening test would automatically determine the disease state. The opposite extreme is achieved when the events {mammogram positive} and {breast cancer} are completely independent. In this case, the probability of breast cancer is the same whether or not the mammogram is positive, and the mammogram would not be useful in screening for breast cancer and should not be used.

These concepts can be quantified in the following way. Let A = {mammogram$^+$}, B = {breast cancer}, and suppose we are interested in the probability of breast cancer (B) given that the mammogram is positive (A). This probability can be written $Pr(A \cap B)/Pr(A)$.

Definition 3.9 The quantity $Pr(A \cap B)/Pr(A)$ is defined as the **conditional probability of *B* given *A***, which is written $Pr(B|A)$.

However, from Section 3.4 we know that, by definition, if two events are independent, then $Pr(A \cap B) = Pr(A) \times Pr(B)$. If both sides are divided by $Pr(A)$, then $Pr(B) = Pr(A \cap B)/Pr(A) = Pr(B|A)$. Similarly, we can show that if A and B are independent events, then $Pr(B|\overline{A}) = Pr(B|A) = Pr(B)$. This relationship leads to the following alternative interpretation of independence in terms of conditional probabilities.

Equation 3.5

(1) If A and B are independent events, then $Pr(B|A) = Pr(B) = Pr(B|\overline{A})$.

(2) If two events A, B are dependent, then $Pr(B|A) \neq Pr(B) \neq Pr(B|\overline{A})$ and $Pr(A \cap B) \neq Pr(A) \times Pr(B)$.

Definition 3.10 The **relative risk (*RR*)** of B given A is

$$Pr(B|A)/Pr(B|\overline{A})$$

Notice that if two events A, B are independent, then the relative risk is 1. If two events A, B are dependent, then the relative risk is different from 1. Heuristically, the more the dependence between events increases, the further the relative risk is from 1.

Example 3.19 **Cancer** Suppose that among 100,000 women with negative mammograms 20 will have breast cancer diagnosed within 2 years, or $Pr(B|\overline{A}) = 20/10^5 = .0002$, whereas 1 woman in 10 with positive mammograms will have breast cancer diagnosed within

2 years, or $Pr(B|A) = .1$. The two events A and B would be highly dependent, because

$$RR = Pr(B|A)/Pr(B|\bar{A}) = .1/.0002 = 500$$

In other words, women with positive mammograms are 500 times as likely to develop breast cancer over the next 2 years than are women with negative mammograms. This is the rationale for using the mammogram as a screening test for breast cancer. If events A and B were independent, then the relative risk would be 1; women with positive or negative mammograms would be equally likely to have breast cancer, and the mammogram would not be useful as a screening test for breast cancer.

Example 3.20 **Sexually Transmitted Disease** Using the data in Example 3.15, find the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a positive diagnosis. What is the conditional probability that doctor B makes a positive diagnosis of syphilis given that doctor A makes a negative diagnosis? What is the relative risk of B^+ given A^+?

Solution

$$Pr(B^+|A^+) = Pr(B^+ \cap A^+)/Pr(A^+) = .08/.1 = .8$$

Thus doctor B will confirm doctor A's positive diagnoses 80% of the time. Similarly,

$$Pr(B^+|A^-) = Pr(B^+ \cap A^-)/Pr(A^-) = Pr(B^+ \cap A^-)/.9$$

We must compute $Pr(B^+ \cap A^-)$. We know that if doctor B diagnoses a patient as positive, then doctor A either does or does not confirm the diagnosis. Thus

$$Pr(B^+) = Pr(B^+ \cap A^+) + Pr(B^+ \cap A^-)$$

because the events $B^+ \cap A^+$ and $B^+ \cap A^-$ are mutually exclusive. If we subtract $Pr(B^+ \cap A^+)$ from both sides of the equation, then

$$Pr(B^+ \cap A^-) = Pr(B^+) - Pr(B^+ \cap A^+) = .17 - .08 = .09$$

Therefore, $Pr(B^+|A^-) = .09/.9 = .1$

Thus, when doctor A diagnoses a patient as negative, doctor B will contradict the diagnosis 10% of the time. The relative risk of the event B^+ given A^+ is

$$Pr(B^+|A^+)/Pr(B^+|A^-) = .8/.1 = 8$$

This indicates that doctor B is 8 times as likely to diagnose a patient as positive when doctor A diagnoses the patient as positive than when doctor A diagnoses the patient as negative. These results quantify the dependence between the two doctors' diagnoses.

REVIEW QUESTIONS 3A

1 What is the frequency definition of probability?

2 What is the difference between independent and dependent events?

REVIEW

3 What are mutually exclusive events?

4 What is the addition law of probability?

5 What is conditional probability? How does it differ from unconditional probability?

6 What is relative risk? How do you interpret it?

Total-Probability Rule

The conditional $\left(Pr(B|A), Pr(B|\overline{A})\right)$ and unconditional $\left(Pr(B)\right)$ probabilities mentioned previously are related in the following way:

Equation 3.6 For any events A and B,

$$Pr(B) = Pr(B|A) \times Pr(A) + Pr(B|\overline{A}) \times Pr(\overline{A})$$

This formula tells us that the unconditional probability of B is the sum of the conditional probability of B given A *times* the unconditional probability of A *plus* the conditional probability of B given A *not* occurring *times* the unconditional probability of A *not* occurring.

To derive this, we note that if the event B occurs, it must occur either with A or without A. Therefore,

$$Pr(B) = Pr(B \cap A) + Pr(B \cap \overline{A})$$

From the definition of conditional probability, we see that

$$Pr(B \cap A) = Pr(A) \times Pr(B|A)$$

and

$$Pr(B \cap \overline{A}) = Pr(\overline{A}) \times Pr(B|\overline{A})$$

By substitution, it follows that

$$Pr(B) = Pr(B|A)Pr(A) + Pr(B|\overline{A})Pr(\overline{A})$$

Example 3.21 **Cancer** Let A and B be defined as in Example 3.19, and suppose that 7% of the general population of women will have a positive mammogram. What is the probability of developing breast cancer over the next 2 years among women in the general population?

Solution

$$\begin{aligned} Pr(B) &= Pr(\text{breast cancer}) \\ &= Pr(\text{breast cancer} \mid \text{mammogram}^+) \times Pr(\text{mammogram}^+) \\ &\quad + Pr(\text{breast cancer} \mid \text{mammogram}^-) \times Pr(\text{mammogram}^-) \\ &= .1(.07) + .0002(.93) = .00719 = 719/10^5 \end{aligned}$$

Thus the unconditional probability of developing breast cancer over the next 2 years in the general population ($719/10^5$) is a weighted average of the conditional probability of developing breast cancer over the next 2 years among women with a positive mammogram (.1) and the conditional probability of developing breast cancer over the next 2 years among women with a negative mammogram ($20/10^5$).

In Equation 3.6 the probability of event B is expressed in terms of two mutually exclusive events A and $\bar{A}$. In many instances the probability of an event B will need to be expressed in terms of more than two mutually exclusive events, denoted by $A_1, A_2, \ldots, A_k$.

Definition 3.11 A set of events $A_1, \ldots, A_k$ is **exhaustive** if at least one of the events must occur.

Assume that events $A_1, \ldots, A_k$ are mutually exclusive and exhaustive; that is, at least one of the events $A_1, \ldots, A_k$ must occur and no two events can occur simultaneously. Thus exactly one of the events $A_1, \ldots, A_k$ must occur.

Equation 3.7 **Total-Probability Rule**

Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. The unconditional probability of B ($Pr(B)$) can then be written as a weighted average of the conditional probabilities of B given $A_i\left(Pr(B|A_i)\right)$ as follows:

$$Pr(B) = \sum_{i=1}^{k} Pr(B|A_i) \times Pr(A_i)$$

To show this, we note that if B occurs, then it must occur together with one and only one of the events, $A_1, \ldots, A_k$. Therefore,

$$Pr(B) = \sum_{i=1}^{k} Pr(B \cap A_i)$$

Also, from the definition of conditional probability,

$$Pr(B \cap A_i) = Pr(A_i) \times Pr(B|A_i)$$

By substitution, we obtain Equation 3.7.

An application of the total-probability rule is given in the following example:

Example 3.22 **Ophthalmology** We are planning a 5-year study of cataract in a population of 5000 people 60 years of age and older. We know from census data that 45% of this population are ages 60–64, 28% are ages 65–69, 20% are ages 70–74, and 7% are age 75 or older. We also know from the Framingham Eye Study that 2.4%, 4.6%, 8.8%, and 15.3% of the people in these respective age groups will develop cataract over the next 5 years [4]. What percentage of our population will develop cataract over the next 5 years, and how many people with cataract does this percentage represent?

Solution Let A_1 = {ages 60–64}, A_2 = {ages 65–69}, A_3 = {ages 70–74}, A_4 = {age 75+}. These events are mutually exclusive and exhaustive, because each person in our population must be in one and only one age group. Furthermore, from the conditions of the problem we know that $Pr(A_1) = .45$, $Pr(A_2) = .28$, $Pr(A_3) = .20$, $Pr(A_4) = .07$, $Pr(B|A_1) = .024$, $Pr(B|A_2) = .046$, $Pr(B|A_3) = .088$, and $Pr(B|A_4) = .153$, where B = {develop cataract in the next 5 years}. Finally, using the total-probability rule,

$$\begin{aligned} Pr(B) &= Pr(B|A_1) \times Pr(A_1) + Pr(B|A_2) \times Pr(A_2) \\ &\quad + Pr(B|A_3) \times Pr(A_3) + Pr(B|A_4) \times Pr(A_4) \\ &= .024(.45) + .046(.28) + .088(.20) + .153(.07) = .052 \end{aligned}$$

Thus 5.2% of our population will develop cataract over the next 5 years, which represents a total of $5000 \times .052 = 260$ people with cataract.

The definition of conditional probability allows the multiplication law of probability to be extended to the case of dependent events.

Equation 3.8 **Generalized Multiplication Law of Probability**

If $A_1, \ldots, A_k$ are an arbitrary set of events, then

$$\begin{aligned} &Pr(A_1 \cap A_2 \cap \cdots \cap A_k) \\ &\quad = Pr(A_1) \times Pr(A_2|A_1) \times Pr(A_3|A_2 \cap A_1) \times \cdots \times Pr(A_k|A_{k-1} \cap \cdots \cap A_2 \cap A_1) \end{aligned}$$

If the events are independent, then the conditional probabilities on the right-hand side of Equation 3.8 reduce to unconditional probabilities, and the generalized multiplication law reduces to the multiplication law for independent events given in Equation 3.2. Equation 3.8 also generalizes the relationship $Pr(A \cap B) = Pr(A) \times Pr(B|A)$ given in Definition 3.9 for two events, to the case of more than two events.

REVIEW QUESTIONS 3B

1 What is the total-probability rule?

2 Suppose the rate of type II diabetes mellitus (DM) among 40- to 59-year-olds is 7% among Caucasians, 10% among African-Americans, 12% among Hispanics, and 5% among Asian-Americans. Suppose the ethnic distribution in Houston, Texas, among 40- to 59-year-olds is 30% Caucasian, 25% African-American, 40% Hispanic, and 5% Asian-American. What is the overall probability of type II DM among 40- to 59-year-olds in Houston?

REVIEW

3.7 Bayes' Rule and Screening Tests

The mammography test given in Example 3.18 illustrates the general concept of the predictive value of a screening test, which can be defined as follows:

Definition 3.12 The **predictive value positive** (PV^+) of a screening test is the probability that a person has a disease given that the test is positive.

$Pr(\text{disease} \mid \text{test}^+)$

The **predictive value negative** (PV^-) of a screening test is the probability that a person does *not* have a disease given that the test is negative.

$Pr(\text{no disease} \mid \text{test}^-)$

Example 3.23 **Cancer** Find the predictive value positive and the predictive value negative for the mammogram given the data in Example 3.19.

Solution We see that $PV^+ = Pr(\text{breast cancer} \mid \text{mammogram}^+) = .1$

whereas $PV^- = Pr(\text{no breast cancer} \mid \text{mammogram}^-)$

$= 1 - Pr(\text{breast cancer} \mid \text{mammogram}^-) = 1 - .0002 = .9998$

Thus, if the mammogram is negative, the woman is virtually certain *not* to develop breast cancer over the next 2 years ($PV^- \approx 1$); whereas if the mammogram is positive, the woman has a 10% chance of developing breast cancer ($PV^+ = .10$).

A symptom or a set of symptoms can also be regarded as a screening test for disease. The higher the predictive value of the screening test or symptoms, the more valuable the test. Ideally, we would like to find a set of symptoms such that both PV^+ and PV^- are 1. Then we could accurately diagnose disease for each patient. However, this is usually impossible.

Clinicians often cannot directly measure the predictive value of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people. These measures are defined as follows:

Definition 3.13 The **sensitivity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is present given that the person has a disease.

Definition 3.14 The **specificity** of a symptom (or set of symptoms or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.

Definition 3.15 A **false negative** is defined as a person who tests out as negative but who is actually positive. A **false positive** is defined as a person who tests out as positive but who is actually negative.

For a symptom to be effective in predicting disease, it is important that both the sensitivity and specificity be high.

Example 3.24 **Cancer** Suppose the disease is lung cancer and the symptom is cigarette smoking. If we assume 90% of people with lung cancer and 30% of people without lung cancer (essentially the entire general population) are smokers, then the sensitivity and

specificity are .9 and .7, respectively. Obviously, cigarette smoking cannot be used by itself as a diagnostic tool for predicting lung cancer, because there will be too many false positives (normal people who are smokers).

Example 3.25 **Cancer** Suppose the disease is breast cancer in women and the symptom is having a family history of breast cancer (either a mother or a sister with breast cancer). If we assume 5% of women with breast cancer have a family history of breast cancer whereas only 2% of women without breast cancer have such a history, then the sensitivity is .05 and the specificity is .98 = (1 – .02). A family history of breast cancer cannot be used by itself to diagnose breast cancer because there will be too many false negatives (women without a family history but who have the disease).

REVIEW QUESTIONS 3C

1 What is the sensitivity and specificity of a screening test?

2 What is the predictive value positive and the predictive value negative of a screening test? How does predictive value differ from sensitivity and specificity?

3 The level of prostate-specific antigen (PSA) in the blood is frequently used as a screening test for prostate cancer. Punglia et al. [5] reported the following data regarding the relationship between a positive PSA test (≥ 4.1 ng/dL) and prostate cancer.

Table 3.2 **PSA and prostate cancer association**

PSA test result	Prostate cancer	Frequency
+	+	92
+	–	27
–	+	46
–	–	72

(a) What is the sensitivity and specificity of the test?

(b) What is the predictive value positive and negative of the test?

Bayes' Rule

Review Question 3C.3 assumes that each PSA^+ and PSA^- participant (or at least a representative sample of PSA^+ and PSA^- participants) is evaluated for the presence of prostate cancer. Thus one can directly evaluate PV^+ and PV^- from the data provided. Instead, in many screening studies, a random sample of cases and controls are obtained. One can estimate sensitivity and specificity from such a design. However, cases are usually oversampled relative to the general population (e.g., there are equal number of cases and controls), so one cannot directly estimate PV^+ and PV^- from the frequency counts available in a typical screening study. Instead, an indirect method known as Bayes' rule is used for this purpose.

The general question is, how can the sensitivity and specificity of a symptom (or set of symptoms), which are quantities a physician can estimate, be used to compute

predictive values, which are quantities a physician needs to make appropriate diagnoses?

Let A = symptom and B = disease. From Definitions 3.12, 3.13, and 3.14, we have

Predictive value positive = $PV^+ = Pr(B|A)$

Predictive value negative = $PV^- = Pr(\overline{B}|\overline{A})$

Sensitivity = $Pr(A|B)$

Specificity = $Pr(\overline{A}|\overline{B})$

Let $Pr(B)$ = probability of disease in the reference population. We wish to compute $Pr(B|A)$ and $Pr(\overline{B}|\overline{A})$ in terms of the other quantities. This relationship is known as Bayes' rule.

Equation 3.9 **Bayes' Rule**

Let A = symptom and B = disease.

$$PV^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})}$$

In words, this can be written as

$$PV^+ = \frac{\text{sensitivity} \times x}{\text{sensitivity} \times x + (1 - \text{specificity}) \times (1 - x)}$$

where $x = Pr(B)$ = prevalence of disease in the reference population. Similarly,

$$PV^- = \frac{\text{specificity} \times (1 - x)}{\text{specificity} \times (1 - x) + (1 - \text{sensitivity}) \times x}$$

To derive this, we have, from the definition of conditional probability,

$$PV^+ = Pr(B|A) = \frac{Pr(B \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B \cap A) = Pr(A|B) \times Pr(B)$$

Finally, from the total-probability rule,

$$Pr(A) = Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})$$

If the expressions for $Pr(B \cap A)$ and $Pr(A)$ are substituted into the equation for PV^+, we obtain

$$PV^+ = Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A|B) \times Pr(B) + Pr(A|\overline{B}) \times Pr(\overline{B})}$$

That is, PV^+ can be expressed as a function of sensitivity, specificity, and probability of disease in the reference population. A similar derivation can be used to obtain PV^-.

Example 3.26 **Hypertension** Suppose 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What are the predictive value positive and predictive value negative of the machine, assuming 20% of the adult population is hypertensive?

Solution The sensitivity = .84 and specificity = 1 − .23 = .77. Thus, from Bayes' rule it follows that

$$PV^+ = (.84)(.2)/[(.84)(.2)+(.23)(.8)]$$
$$= .168/.352 = .48$$

Similarly, $$PV^- = (.77)(.8)/[(.77)(.8)+(.16)(.2)]$$
$$= .616/.648 = .95$$

Thus a negative result from the machine is very predictive, because we are 95% sure a person with a negative result from the machine is normotensive. However, a positive result is not very predictive, because we are only 48% sure a person with a positive result from the machine is hypertensive.

Example 3.26 considered only two possible disease states: hypertensive and normotensive. In clinical medicine there are often more than two possible disease states. We would like to be able to predict the most likely disease state given a specific symptom (or set of symptoms). Let's assume the probability of having these symptoms among people in a given disease state is known from clinical experience, as is the probability of each disease state in the reference population. This leads us to the generalized Bayes' rule:

Equation 3.10 **Generalized Bayes' Rule**

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive disease states; that is, at least one disease state must occur and no two disease states can occur at the same time. Let A represent the presence of a symptom or set of symptoms. Then

$$Pr(B_i|A) = Pr(A|B_i) \times Pr(B_i) \Big/ \left[\sum_{j=1}^{k} Pr(A|B_j) \times Pr(B_j) \right]$$

This result is obtained similarly to the result of Bayes' rule for two disease states in Equation 3.9. Specifically, from the definition of conditional probability, note that

$$Pr(B_i|A) = \frac{Pr(B_i \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B_i \cap A) = Pr(A|B_i) \times Pr(B_i)$$

From the total-probability rule,

$$Pr(A) = Pr(A|B_1) \times Pr(B_1) + \cdots + Pr(A|B_k) \times Pr(B_k)$$

If the expressions for $Pr(B_i \cap A)$ and $Pr(A)$ are substituted, we obtain

$$Pr(B_i|A) = \frac{Pr(A|B_i) \times Pr(B_i)}{\sum_{j=1}^{k} Pr(A|B_j) \times Pr(B_j)}$$

Example 3.27 **Pulmonary Disease** Suppose a 60-year-old man who has never smoked cigarettes presents to a physician with symptoms of a chronic cough and occasional breathlessness. The physician becomes concerned and orders the patient admitted to the hospital for a lung biopsy. Suppose the results of the lung biopsy are consistent either with lung cancer or with sarcoidosis, a fairly common, nonfatal lung disease. In this case

Symptoms A = {chronic cough, results of lung biopsy}
Disease state B_1 = normal
B_2 = lung cancer
B_3 = sarcoidosis

Suppose that $Pr(A|B_1) = .001$ $Pr(A|B_2) = .9$ $Pr(A|B_3) = .9$

and that in 60-year-old, never-smoking men

$$Pr(B_1) = .99 \quad Pr(B_2) = .001 \quad Pr(B_3) = .009$$

The first set of probabilities $Pr(A|B_i)$ could be obtained from clinical experience with the previous diseases, whereas the latter set of probabilities $Pr(B_i)$ would have to be obtained from age–sex–smoking specific prevalence rates for the diseases in question. The interesting question now is, What are the probabilities $Pr(B_i|A)$ of the three disease states given the previous symptoms?

Solution Bayes' rule can be used to answer this question. Specifically,

$$\begin{aligned} Pr(B_1|A) &= Pr(A|B_1) \times Pr(B_1) \Big/ \left[\sum_{j=1}^{3} Pr(A|B_j) \times Pr(B_j)\right] \\ &= .001(.99)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00099/.00999 = .099 \end{aligned}$$

$$\begin{aligned} Pr(B_2|A) &= .9(.001)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00090/.00999 = .090 \end{aligned}$$

$$\begin{aligned} Pr(B_3|A) &= .9(.009)/[.001(.99) + .9(.001) + .9(.009)] \\ &= .00810/.00999 = .811 \end{aligned}$$

Thus, although the unconditional probability of sarcoidosis is very low (.009), the conditional probability of the disease given these symptoms and this age–sex–

smoking group is .811. Also, although the symptoms are consistent with both lung cancer and sarcoidosis, the latter is much more likely among patients in this age–sex–smoking group.

Example 3.28 **Pulmonary Disease** Now suppose the patient in Example 3.27 smoked 2 packs of cigarettes per day for 40 years. Then assume $Pr(B_1) = .98$, $Pr(B_2) = .015$, $Pr(B_3) = .005$ in this type of person. What are the probabilities of the three disease states for this type of patient, given these symptoms?

Solution

$$Pr(B_1|A) = .001(.98)/[.001(.98) + .9(.015) + .9(.005)]$$
$$= .00098/.01898 = .052$$
$$Pr(B_2|A) = .9(.015)/.01898 = .01350/.01898 = .711$$
$$Pr(B_3|A) = .9(.005)/.01898 = .237$$

Thus in this type of patient lung cancer is the most likely diagnosis.

REVIEW QUESTIONS 3D

1 What is Bayes' rule? How is it used?

2 What is the generalized Bayes' rule?

3 Refer to Review Question 3B.2. Suppose a 40- to 59-year-old person in Houston, Texas, has type II diabetes mellitus (DM). What is the probability that this person is African-American? Hispanic? Caucasian? Asian-American? *Hint:* Use the generalized Bayes' rule.

4 Answer Review Question 3D.3 for a nondiabetic 40- to 59-year-old person in Houston.

REVIEW

3.8 Bayesian Inference

The definition of probability given in Definition 3.1 is sometimes called the **frequency definition of probability.** It forms the basis for the frequentist method of inference, which is the main approach to statistical inference featured in this book and used in statistical practice. However, Bayesian inference is an alternative definition of probability and inference, espoused by a vocal minority of statisticians. The Bayesian school of inference rejects the idea of the frequency definition of probability, considering this idea a theoretical concept that can never be realized in practice. Instead, Bayesians conceive of two types of probability: a prior probability and a posterior probability.

Definition 3.16 The **prior probability** of an event is the best guess by the observer of an event's probability in the absence of data. This prior probability may be a single number, or it may be a range of likely values for the probability, perhaps with weights attached to each possible value.

Example 3.29 **Hypertension** What is the prior probability of hypertension in Example 3.26?

Solution The prior probability of hypertension in the absence of additional data is .20, because 20% of the adult population is hypertensive.

Definition 3.17 The **posterior probability** of an event is the probability of an event after collecting some empirical data. It is obtained by integrating information from the prior probability with additional data related to the event in question.

Example 3.30 **Hypertension** What is the posterior probability of hypertension given that an automated blood-pressure machine has classified a person as hypertensive?

Solution If we refer to Example 3.26, and let the event {true hypertensive} be denoted by B and the event {classified as hypertensive by an automated blood-pressure machine} be denoted by A, we see that the posterior probability is given by the predictive value positive $(PV^+) = Pr(B|A) = .48$.

Example 3.31 **Hypertension** What is the posterior probability of hypertension given that an automated blood-pressure machine has classified a person as normotensive?

Solution The posterior probability $= Pr(B|\overline{A}) = 1 - Pr(\overline{B}|\overline{A}) = 1 - PV^- = .05$. Thus, the initial prior probability of 20% has been integrated with the automated blood-pressure machine data to yield posterior probabilities of .48 and .05, for people who are classified as hypertensive and normotensive by the automated blood-pressure machine, respectively.

The main problem with Bayesian inference lies in specifying the prior probability. Two different people may provide different prior probabilities for an event and may reach different conclusions (obtain different posterior probabilities), even with the same data. However, in some cases the prior probability is well defined. Also, sufficient data diminishes the impact of the prior probability on the posterior inference. Although most methodology in this book is based on the frequentist approach to inference, we use Bayesian inference in several additional examples in later chapters to provide further insight into data analysis.

3.9 ROC Curves

In some instances, a test provides several categories of response rather than simply test positive or test negative. In other instances, the results of the test are reported as a continuous variable. In either case, designation of a cutoff point for distinguishing test positive versus test negative is arbitrary.

Example 3.32 **Radiology** The following data, provided by Hanley and McNeil [6], are ratings of computed tomographic (CT) images by a single radiologist in a sample of 109 subjects with possible neurologic problems. The true disease status is also known for

Table 3.3 **Ratings of 109 CT images by a single radiologist**

	CT rating by radiologist					
True disease status	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)	Total
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

each of these subjects. The data are presented in Table 3.3. How can we quantify the diagnostic accuracy of the test?

Unlike previous examples, this one has no obvious cutoff point to use for designating a subject as positive based on the CT scan. For example, if we designate a subject as test positive if he or she is either probably abnormal or definitely abnormal (a rating of 4 or 5, or 4+), then the sensitivity of the test is $(11 + 33)/51 = 44/51 = .86$, whereas the specificity is $(33 + 6 + 6)/58 = 45/58 = .78$. In Table 3.4, we compute the sensitivity and specificity of the radiologist ratings according to different criteria for test positive.

Table 3.4 **Sensitivity and specificity of the radiologist ratings according to different test-positive criteria based on the data in Table 3.3**

Test-positive criteria	Sensitivity	Specificity
1+	1.0	0
2+	.94	.57
3+	.90	.67
4+	.86	.78
5+	.65	.97
6+	0	1.0

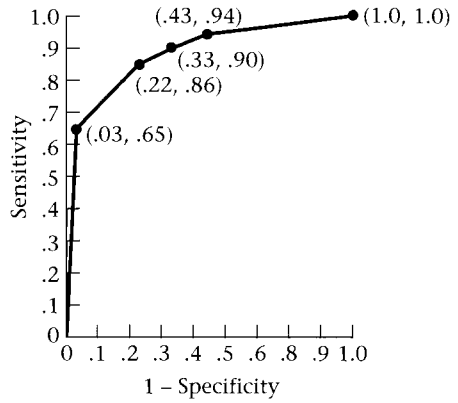
To display these data, we construct a receiver operating characteristic (ROC) curve.

Definition 3.18 A **receiver operating characteristic (ROC) curve** is a plot of the sensitivity versus (1 – specificity) of a screening test, where the different points on the curve correspond to different cutoff points used to designate test positive.

Example 3.33 **Radiology** Construct an ROC curve based on the data in Table 3.4.

Solution We plot sensitivity versus (1 – specificity) using the data in Table 3.4. The plot is shown in Figure 3.7.

Figure 3.7 ROC curve for the data in Table 3.4*



*Each point represents (1 – specificity, sensitivity) for different test-positive criteria.

The area under the ROC curve is a reasonable summary of the overall diagnostic accuracy of the test. It can be shown [6] that this area, when calculated by the trapezoidal rule, corresponds to the probability that for a randomly selected pair of normal and abnormal subjects, the radiologist will correctly identify the normal subject given the CT ratings. It is assumed that for untied ratings the radiologist designates the subject with the lower test score as normal, and the subject with the higher test score as abnormal. For tied ratings, it is assumed the radiologist randomly chooses one patient as normal and the other as abnormal.

Example 3.34 **Radiology** Calculate the area under the ROC curve in Figure 3.7, and interpret what it means.

Solution The area under the ROC curve, when evaluated by the trapezoidal rule, is given by

$$.5(.94 + 1.0)(.57) + .5\,(.90 + .94)(.10) + .5(.86 + .90)(.11) + .5(.65 + .86)(.19) + .5(0 + .65)(.03) = .89$$

This means the radiologist has an 89% probability of correctly distinguishing a normal from an abnormal subject based on the relative ordering of their CT ratings. For normal and abnormal subjects with the same ratings, it is assumed the radiologist selects one of the two subjects at random.

In general, of two screening tests for the same disease, the test with the higher area under its ROC curve is considered the better test, unless some particular level of sensitivity or specificity is especially important in comparing the two tests.

3.10 Prevalence and Incidence

In clinical medicine, the terms *prevalence* and *incidence* denote probabilities in a special context and are used frequently in this text.

Definition 3.19 The **prevalence** of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease. It is obtained by dividing the number of people who currently have the disease by the number of people in the study population.

Example 3.35 **Hypertension** The prevalence of hypertension in 1974 among all people 17 years of age and older was reported to be 15.7%, as assessed by a government study [7]. It was computed by dividing the number of people who had elevated blood pressure and were 17 years of age and older (22,626) by the total number of people 17 years of age and older in the study population (144,380).

Definition 3.20 The **cumulative incidence** of a disease is the probability that a person with no prior disease will develop a new case of the disease over some specified time period.

In Chapter 14 we distinguish between *cumulative incidence,* which is defined over a long period of time, and *incidence density,* which is defined over a very short (or instantaneous) period of time. For simplicity, before Chapter 14 we use the abbreviated term *incidence* to denote *cumulative incidence.*

Example 3.36 **Cancer** The cumulative-incidence rate of breast cancer in 40- to 44-year-old Connecticut women over the time period January 1, 1970, through December 31, 1970, was approximately 1 per 1000 [2]. This rate means that on January 1, 1970, about 1 woman in 1000 of the 40- to 44-year-old women who had never had breast cancer would have developed a new case of breast cancer by December 31, 1970.

REVIEW QUESTIONS 3E

1 Suppose 25 students are in a class, 5 of them currently suffering from hay fever. Is the proportion 5 out of 25 (20%) a measure of prevalence, of incidence, or of neither?

2 Suppose 50 HIV-positive men are identified. Five of the men develop AIDS over the next 2 years. Is the proportion 5 out of 50 (10%) a measure of prevalence, of incidence, or of neither?

REVIEW

3.11 Summary

In this chapter, probabilities and how to work with them using the addition and multiplication laws were discussed. An important distinction was made between independent events, which are unrelated to each other, and dependent events, which are related to each other. The general concepts of conditional probability and relative risk were introduced to quantify the dependence between two events. These ideas were then applied to the special area of screening populations for disease. In particular, the notions of sensitivity, specificity, and predictive value, which are used to define the accuracy of screening tests, were developed as applications of conditional probability. We also used an ROC curve to extend the concepts of sensitivity

and specificity when the designation of the cutoff point for test positive versus test negative is arbitrary.

On some occasions, only sensitivities and specificities are available and we wish to compute the predictive value of screening tests. This task can be accomplished using Bayes' rule. The use of Bayes' rule in the context of screening tests is a special case of Bayesian inference. In Bayesian inference, we specify a prior probability for an event, which after data are collected is then modified to a posterior probability. Finally, prevalence and incidence, which are probabilistic parameters that are often used to describe the magnitude of disease in a population, were defined.

In the next two chapters, these general principles of probability are applied to derive some of the important probabilistic models often used in biomedical research, including the binomial, Poisson, and normal models. These models will eventually be used to test hypotheses about data.

PROBLEMS

Consider a family with a mother, father, and two children. Let A_1 = {mother has influenza}, A_2 = {father has influenza}, A_3 = {first child has influenza}, A_4= {second child has influenza}, B = {at least one child has influenza}, C = {at least one parent has influenza}, D = {at least one person in the family has influenza}.

*3.1 What does $A_1 \cup A_2$ mean?

*3.2 What does $A_1 \cap A_2$ mean?

*3.3 Are A_3 and A_4 mutually exclusive?

*3.4 What does $A_3 \cup B$ mean?

*3.5 What does $A_3 \cap B$ mean?

*3.6 Express C in terms of A_1, A_2, A_3, A_4.

*3.7 Express D in terms of B and C.

*3.8 What does $\overline{A}_1$ mean?

*3.9 What does $\overline{A}_2$ mean?

*3.10 Represent $\overline{C}$ in terms of A_1, A_2, A_3, A_4.

*3.11 Represent $\overline{D}$ in terms of B and C.

Suppose an influenza epidemic strikes a city. In 10% of families the mother has influenza; in 10% of families the father has influenza; and in 2% of families both the mother and father have influenza.

3.12 Are events A_1 = {mother has influenza} and A_2 = {father has influenza} independent?

Suppose there is a 20% chance each child will get influenza, whereas in 10% of two-child families both children get the disease.

3.13 What is the probability that at least one child will get influenza?

Hypertension

Multiple drugs are often used in treating hypertension. Suppose 10% of patients taking antihypertensive agent A experience gastrointestinal (GI) side effects, whereas 20% of patients taking antihypertensive agent B experience such side effects.

3.14 If the side effects of the two agents are assumed to be independent events, then what is the probability that a patient taking the two agents simultaneously will experience GI side effects from at least one of the two agents?

Look at Problem 3.12.

3.15 What is the conditional probability that the father has influenza given that the mother has influenza?

3.16 What is the conditional probability that the father has influenza given that the mother does not have influenza?

Mental Health

Estimates of the prevalence of Alzheimer's disease have recently been provided by Pfeffer et al. [8]. The estimates are given in Table 3.5.

Suppose an unrelated 77-year-old man, 76-year-old woman, and 82-year-old woman are selected from a community.

3.17 What is the probability that all three of these individuals have Alzheimer's disease?

3.18 What is the probability that at least one of the women has Alzheimer's disease?

3.19 What is the probability that at least one of the three people has Alzheimer's disease?

Table 3.5 Prevalence of Alzheimer's disease (cases per 100 population)

Age group	Males	Females
65–69	1.6	0.0
70–74	0.0	2.2
75–79	4.9	2.3
80–84	8.6	7.8
85+	35.0	27.9

Table 3.6 Age–sex distribution of retirement community

	Male (%)[a]	Female (%)
65–69	5	10
70–74	9	17
75–79	11	18
80–84	8	12
85+	4	6

[a]Percentage of total population.

3.20 What is the probability that exactly one of the three people has Alzheimer's disease?

3.21 Suppose we know one of the three people has Alzheimer's disease, but we don't know which one. What is the conditional probability that the affected person is a woman?

3.22 Suppose we know two of the three people have Alzheimer's disease. What is the conditional probability that they are both women?

3.23 Suppose we know two of the three people have Alzheimer's disease. What is the conditional probability that they are both less than 80 years old?

Suppose the probability that both members of a married couple will have the disease, where each member is 75–79 years old, is .0015.

3.24 What is the conditional probability that the man will be affected given that the woman is affected? How does this value compare to the prevalence in Table 3.5? Why should it be the same (or different)?

3.25 What is the conditional probability that the woman will be affected given that the man is affected? How does this value compare to the prevalence in Table 3.5? Why should it be the same (or different)?

3.26 What is the probability that at least one member of the couple is affected?

Suppose a study of Alzheimer's disease is proposed in a retirement community with people 65+ years of age, where the age–sex distribution is as shown in Table 3.6.

3.27 What is the expected overall prevalence of Alzheimer's disease in the community if the prevalence estimates in Table 3.5 for specific age–sex groups hold?

3.28 If 1000 people 65+ years old are in the community, then what is the expected number of cases of Alzheimer's disease in the community?

Occupational Health

A study is conducted among men 50–69 years old working in a chemical plant. We are interested in comparing the mortality experience of the workers in the plant with national mortality rates. Suppose that of the 500 workers in this age group in the plant, 35% are 50–54, 30% are 55–59, 20% are 60–64, and 15% are 65–69.

***3.29** If the annual national mortality rates are 0.9% in 50- to 54-year-old men, 1.4% in 55- to 59-year-old men, 2.2% in 60- to 64-year-old men, and 3.3% in 65- to 69-year-old men, then what is the projected annual mortality rate in the plant as a whole?

The SMR (standardized mortality ratio) is often used in occupational studies as a measure of risk. It is defined as 100% *times* the observed number of events in the exposed group *divided by* the expected number of events in the exposed group (based on some reference population).

***3.30** If 15 deaths are observed over 1 year among the 500 workers, what is the SMR?

Genetics

Suppose that a disease is inherited via a **dominant mode of inheritance** and that one of two parents is affected with the disease whereas one is not. The implications of this mode of inheritance are that the probability is 1/2 that any particular offspring will get the disease.

3.31 What is the probability that in a family with two children, both siblings are affected?

3.32 What is the probability that exactly one sibling is affected?

3.33 What is the probability that neither sibling will be affected?

3.34 Suppose the older child is affected. What is the probability that the younger child will be affected?

3.35 If A, B are two events such that A = {older child is affected}, B = {younger child is affected}, then are the events A, B independent?

Suppose that a disease is inherited via an **autosomal recessive mode of inheritance.** The implications of this mode of inheritance are that the children in a family each have a probability of 1/4 of inheriting the disease.

3.36 What is the probability that in a family with two children, both siblings are affected?

3.37 What is the probability that exactly one sibling is affected?

3.38 What is the probability that neither sibling is affected?

Suppose that a disease is inherited via a **sex-linked mode of inheritance.** The implications of this mode of inheritance are that each male offspring has a 50% chance of inheriting the disease, whereas the female offspring have no chance of getting the disease.

3.39 In a family with one male and one female offspring, what is the probability that both siblings are affected?

3.40 What is the probability that exactly one sibling is affected?

3.41 What is the probability that neither sibling is affected?

3.42 Answer Problem 3.39 for families with two male siblings.

3.43 Answer Problem 3.40 for families with two male siblings.

3.44 Answer Problem 3.41 for families with two male siblings.

Suppose that in a family with two male siblings, both siblings are affected with a genetically inherited disease. Suppose also that, although the genetic history of the family is unknown, only a dominant, recessive, or sex-linked mode of inheritance is possible.

3.45 Assume that the dominant, recessive, and sex-linked modes of inheritance follow the probability laws given in Problems 3.31, 3.36, and 3.39 and that, without prior knowledge about the family in question, each is equally likely. What is the posterior probability of each mode of inheritance in this family?

3.46 Answer Problem 3.45 for a family with two male siblings where only one sibling is affected.

3.47 Answer Problem 3.45 for a family with one male and one female sibling where both siblings are affected.

3.48 Answer Problem 3.47 where only the male sibling is affected.

Obstetrics

The following data are derived from the 1973 Final Natality Statistics Report issued by the National Center for Health Statistics [9]. These data are pertinent to livebirths only.

Suppose that infants are classified as low birthweight if they have a birthweight ≤ 2500 g and as normal birthweight if they have a birthweight ≥ 2501 g. Suppose that infants are also classified by length of gestation in the following four categories: <20 weeks, 20–27 weeks, 28–36 weeks, >36 weeks. Assume the probabilities of the different periods of gestation are as given in Table 3.7.

Table 3.7 Distribution of length of gestation

Length of gestation	Probability
<20 weeks	.0004
20–27 weeks	.0059
28–36 weeks	.0855
>36 weeks	.9082

Also assume that the probability of low birthweight given that length of gestation is <20 weeks is .540, the probability of low birthweight given that length of gestation is 20–27 weeks is .813, the probability of low birthweight given that length of gestation is 28–36 weeks is .379, and the probability of low birthweight given that length of gestation is >36 weeks is .035.

***3.49** What is the probability of having a low birthweight infant?

3.50 Show that the events {length of gestation ≤ 27 weeks} and {low birthweight} are not independent.

***3.51** What is the probability of having a length of gestation ≤ 36 weeks given that a child is low birthweight?

Pulmonary Disease

A 1974 paper by Colley et al. looked at the relationship between parental smoking and the incidence of pneumonia and/or bronchitis in children in the first year of life [10]. One important finding of the paper was that 7.8% of children with nonsmoking parents had episodes of pneumonia and/or bronchitis in the first year of life, whereas, respectively, 11.4% of children with one smoking parent and 17.6% of children with two smoking parents had such an episode. Suppose that in the general population both parents are smokers in 40% of households, one parent smokes in 25% of households, and neither parent smokes in 35% of households.

3.52 What percentage of children in the general population will have pneumonia and/or bronchitis in the first year of life?

A group of families in which both parents smoke at the time of the first prenatal visit decide, after counseling by the nurse practitioner, to give up smoking. Suppose that in 10% of these families both parents resume smoking and in 30% of these families one parent resumes smoking. In the remainder of the families both parents have not resumed smoking at the time the child is born. Assume also that the smoking status of the parents at time of birth is maintained during the child's first year of life.

3.53 What is the probability of pneumonia and/or bronchitis in children from families in this group?

3.54 Suppose in 1000 families both parents smoke. In the absence of any counseling, how many of their children would be expected to have pneumonia or bronchitis in the first year of life?

3.55 Answer Problem 3.54 if the parents get counseling for smoking cessation.

3.56 What percentage of the cases that would have accrued without counseling were prevented by counseling for smoking cessation?

Pulmonary Disease

The familial aggregation of respiratory disease is a well-established clinical phenomenon. However, whether this aggregation is due to genetic or environmental factors or both is somewhat controversial. An investigator wishes to study a particular environmental factor, namely the relationship of cigarette-smoking habits in the parents to the presence or absence of asthma in their oldest child living in the household in the 5- to 9-year-old age range (referred to below as their offspring). Suppose the investigator finds that (1) if both the mother and father are current smokers, then the probability of their offspring having asthma is .15; (2) if the mother is a current smoker and the father is not, then the probability of their offspring having asthma is .13; (3) if the father is a current smoker and the mother is not, then the probability of their offspring having asthma is .05; (4) if neither parent is a current smoker, then the probability of their offspring having asthma is .04.

***3.57** Suppose the smoking habits of the parents are independent and the probability that the mother is a current smoker is .4, whereas the probability that the father is a current smoker is .5. What is the probability that both the father and mother are current smokers?

***3.58** Consider the subgroup of families where the mother is not a current smoker. What is the probability that the father is a current smoker among such families? How does this probability differ from that calculated in Problem 3.57?

Suppose, alternatively, that if the father is a current smoker, then the probability that the mother is a current smoker is .6; whereas if the father is not a current smoker, then the probability that the mother is a current smoker is .2. Also assume that statements 1, 2, 3, and 4, above hold.

***3.59** If the probability that the father is a current smoker is .5, what is the probability that the father is a current smoker *and* that the mother is not a current smoker?

***3.60** Are the current smoking habits of the father and the mother independent? Why or why not?

***3.61** Under the assumptions in Problems 3.59 and 3.60, find the unconditional probability that the offspring will have asthma.

***3.62** Suppose a child has asthma. What is the posterior probability that the father is a current smoker?

***3.63** What is the posterior probability that the mother is a current smoker if the child has asthma?

***3.64** Answer Problem 3.62 if the child does not have asthma.

***3.65** Answer Problem 3.63 if the child does not have asthma.

***3.66** Are the child's asthma status and the father's smoking status independent? Why or why not?

***3.67** Are the child's asthma status and the mother's smoking status independent? Why or why not?

Pulmonary Disease

Smoking cessation is an important dimension in public-health programs aimed at preventing cancer and heart and lung diseases. For this purpose, data were accumulated starting in 1962 on a group of currently smoking men as part of the Normative Aging Study, a longitudinal study of the Veterans Administration in Boston. No interventions were attempted on this group of men, but the data in Table 3.8 were obtained as to annual quitting rates among initially healthy men who remained healthy during the entire period [11].

Note that the quitting rates increased from 1967 to 1970, which was around the time of the first U.S. Surgeon General's report on cigarette smoking.

3.68 Suppose a man was a light smoker on January 1, 1962. What is the probability that he quit smoking by the end of 1975? (Assume he remained a light smoker until just prior to quitting and remained a quitter until 1975.)

3.69 Answer Problem 3.68 for a heavy smoker on January 1, 1962 (assuming he remained a heavy smoker until just prior to quitting and remained a quitter until 1975).

Table 3.8 Annual quitting rates of men who smoked, from the Normative Aging Study, 1962–1975

Time period	Light smokers (≤ one pack per day) average annual quitting rate per 100 people	Heavy smokers (> one pack per day) average annual quitting rate per 100 people
1962–1966	3.1	2.0
1967–1970	7.1	5.0
1971–1975	4.7	4.1

Pulmonary Disease

Research into cigarette-smoking habits, smoking prevention, and cessation programs necessitates accurate measurement of smoking behavior. However, decreasing social acceptability of smoking appears to cause significant underreporting. Chemical markers for cigarette use can provide objective indicators of smoking behavior. One widely used noninvasive marker is the level of saliva thiocyanate (SCN). In a Minneapolis school district, 1332 students in eighth grade (ages 12–14) participated in a study [12] whereby they

(1) Viewed a film illustrating how recent cigarette use could be readily detected from small samples of saliva

(2) Provided a personal sample of saliva thiocyanate

(3) Provided a self-report of the number of cigarettes smoked per week

The results are given in Table 3.9.

Table 3.9 Relationship between saliva thiocyanate levels (SCN) and self-reported cigarettes smoked per week

Self-reported cigarettes smoked in last week	Number of students	Percent with SCN ≥ 100 μg/mL
None	1163	3.3
1–4	70	4.3
5–14	30	6.7
15–24	27	29.6
25–44	19	36.8
45+	23	65.2

Source: Reprinted with permission from the *American Journal of Public Health*, 71(12), 1320, 1981.

Suppose the self-reports are completely accurate and are representative of the amount eighth-grade students smoke in the general community. We are considering using an SCN level ≥ 100 μg/mL as a test criterion for identifying cigarette smokers. Regard a student as positive if he or she smokes one or more cigarettes per week.

***3.70** What is the sensitivity of the test for light-smoking students (students who smoke ≤ 14 cigarettes per week)?

***3.71** What is the sensitivity of the test for moderate-smoking students (students who smoke 15–44 cigarettes per week)?

***3.72** What is the sensitivity of the test for heavy-smoking students (students who smoke ≥ 45 cigarettes per week)?

***3.73** What is the specificity of the test?

***3.74** What is the predictive value positive of the test?

***3.75** What is the predictive value negative of the test?

Suppose we regard the self-reports of all students who report some cigarette consumption as valid but estimate that 20% of students who report no cigarette consumption actually smoke 1–4 cigarettes per week and an additional 10% smoke 5–14 cigarettes per week.

***3.76** If we assume the percentage of students with SCN ≥ 100 μg/mL in these two subgroups is the same as in those who truly report 1–4 and 5–14 cigarettes per week, then compute the specificity under these assumptions.

***3.77** Compute the predictive value negative (PV^-) under these altered assumptions. How does the true PV^- using a screening criteria of SCN ≥ 100 μg/mL for identifying smokers compare with the PV^- based on self-reports obtained in Problem 3.75?

Hypertension

Laboratory measures of cardiovascular reactivity are receiving increasing attention. Much of the expanded interest is based on the belief that these measures, obtained under challenge from physical and psychological stressors, may yield a more biologically meaningful index of cardiovascular function than more traditional static measures. Typically, measurement of cardiovascular reactivity involves the use of an automated blood-pressure monitor to examine the changes in blood pressure before and after a stimulating experience (such as playing a video game). For this purpose, BP measurements were made with the Vita-Stat BP machine both before and after playing a video game. Similar measurements were obtained using manual methods for measuring blood pressure. A person was classified as a "reactor" if his or her diastolic blood pressure (DBP) increased by 10 mm Hg or more after playing the game and as a nonreactor otherwise. The results are given in Table 3.10.

Table 3.10 Classification of cardiovascular reactivity using an automated and a manual sphygmomanometer

	ΔDBP, manual	
ΔDBP, automated	<10	≥10
<10	51	7
≥10	15	6

3.78 If the manual measurements are regarded as the "true" measure of reactivity, then what is the sensitivity of automated DBP measurements?

3.79 What is the specificity of automated DBP measurements?

3.80 If the population tested is representative of the general population, then what are the predictive values positive and negative using this test?

Otolaryngology

The data set in Table 3.11 is based on 214 children with acute otitis media (otitis media with effusion, or OME) who participated in a randomized clinical trial [13]. Each child had OME at the beginning of the study in either one (unilateral cases) or both (bilateral cases) ears. Each child was randomly assigned to receive a 14-day course of one of two antibiotics, either cefaclor (CEF) or amoxicillin (AMO). The data here concern the 203 children whose middle-ear status was determined at a 14-day follow-up visit. The data in Table 3.11 are presented in Data Set EAR.DAT (on the CD-ROM).

3.81 Does there seem to be any difference in the effect of the antibiotics on clearance of otitis media? Express your results in terms of relative risk. Consider separate analyses for unilateral and bilateral cases. Also consider an analysis combining the two types of cases.

Table 3.11 Format for EAR.DAT

Column	Variable	Format or code
1–3	ID	
5	Clearance by 14 days	1 = yes/0 = no
7	Antibiotic	1 = CEF/2 = AMO
9	Age	1 = < 2 yrs/2 = 2–5 yrs/3 = 6+ yrs
11	Ear	1 = 1st ear/2 = 2nd ear

3.82 The investigators recorded the age of the children because they felt this might be an important factor in determining outcome. Were they right? Try to express your results in terms of relative risk.

3.83 While controlling for age, propose an analysis comparing the effectiveness of the two antibiotics. Express your results in terms of relative risk.

3.84 Another issue in this trial is the possible dependence between ears for the bilateral cases. Comment on this issue based on the data collected.

The concept of a **randomized clinical trial** is discussed more completely in Chapter 6. The analysis of **contingency-table data** is studied in Chapters 10 and 13, where many of the formal methods for analyzing this type of data are discussed.

Cardiovascular Disease

A group from the University of Utah set up an experiment to use Bayes' rule to help make clinical diagnoses [14]. In particular, a detailed medical history questionnaire and electrocardiogram were administered to each patient referred to a cardiovascular laboratory and suspected of having congenital heart disease. From the experience of this laboratory and from estimates based on other published data, two sets of probabilities were generated:

(1) The unconditional probability of each of several disease states (see prevalence column in Table 3.12).

(2) The conditional probability of specific symptoms given specific disease states (see the rest of Table 3.12).

Thus, the probability that a person has chest pain given that he or she is normal is .05. Similarly, the proportion of persons with isolated pulmonary hypertension is .020. A subset of the data is given in Table 3.12.

Assume that these diagnoses are the only ones possible and that a patient can have one and only one diagnosis.

3.85 What is the probability of having a symptom of chest pain given that you have isolated pulmonary hypertension?

3.86 What is the probability of being more than 20 years old in this clinic?

3.87 What is the prior probability of being diagnosed as normal in this clinic (i.e., in the absence of any data regarding symptoms)?

3.88 Suppose we assume the probabilities of any set of symptoms are independent given a specific diagnosis (e.g., the probability of being >20 years old and having both chest pain and mild cyanosis given that one is normal is $.50 \times .05 \times .01 = .00025$). What is the posterior probability

Table 3.12 Prevalence of symptoms and diagnoses for patients suspected of having congenital heart disease

		Symptoms						
Diagnosis	Prevalence	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y_1	.155	.49	.50	.01	.10	.05	.05	.01
Y_2	.126	.50	.50	.02	.50	.02	.40	.70
Y_3	.084	.55	.05	.25	.90	.05	.10	.95
Y_4	.020	.45	.45	.01	.95	.10	.10	.95
Y_5	.098	.10	.00	.20	.70	.01	.05	.40
Y_6	.391	.70	.15	.01	.30	.01	.15	.30
Y_7	.126	.60	.10	.30	.70	.10	.20	.70

Y_1 = normal
Y_2 = atrial septal defect without pulmonary stenosis or pulmonary hypertension[a]
Y_3 = ventricular septal defect with valvular pulmonary stenosis
Y_4 = isolated pulmonary hypertension[a]
Y_5 = transposed great vessels
Y_6 = ventricular septal defect without pulmonary hypertension[a]
Y_7 = ventricular septal defect with pulmonary hypertension[a]
X_1 = age 1–20 years old
X_2 = age >20 years old
X_3 = mild cyanosis
X_4 = easy fatigue
X_5 = chest pain
X_6 = repeated respiratory infections
X_7 = EKG axis more than 110°

[a]Pulmonary hypertension is defined as pulmonary artery pressure ≥ systematic arterial pressure.
Source: Reprinted with permission of the American Medical Association from *JAMA, 177*(3), 177–183, 1961. Copyright 1961, American Medical Association.

of being diagnosed as normal given that you have the following symptoms: (1) age 1–20 years, (2) repeated respiratory infections, and (3) easy fatigue?

3.89 The data set in Table 3.12 is in HEART.DAT on the CD-ROM. Use Excel (or some other computer package) to compute the posterior probability of each possible diagnosis given that a patient has all the following: (1) mild cyanosis, (2) age >20 years, (3) EKG axis more than 110°. Construct a bar graph giving the prior probability and the posterior probability for each diagnosis. If a patient has the preceding symptoms, what is the most likely diagnosis?

Suppose the symptom of an EKG axis of more than 110° is used as a screening criterion for diagnosing atrial septal defect without pulmonary stenosis or pulmonary hypertension.

3.90 What is the sensitivity of this test?

3.91 What is the specificity of this test?

3.92 Suppose we want to use another symptom in addition to EKG axis more than 110° to diagnose atrial septal defect without pulmonary stenosis or pulmonary hypertension. If we use the symptoms of age 1–20 years and EKG axis more than 110°, then what is the predictive value positive?

3.93 Suppose we want to use two symptoms to diagnose atrial septal defect without pulmonary stenosis or pulmonary hypertension (not necessarily including EKG axis more than 110°). Which two symptoms can be used to maximize the predictive value positive? (Use a computer to answer this question.)

3.94 Answer Problem 3.93 using three symptoms rather than two. (Use a computer to answer this question.)

Gynecology

A drug company is developing a new pregnancy-test kit for use on an outpatient basis. The company uses the pregnancy test on 100 women who are known to be pregnant, of whom 95 are positive using the test. The company uses the pregnancy test on 100 other women who are known to *not* be pregnant, of whom 99 are negative using the test.

***3.95** What is the sensitivity of the test?

***3.96** What is the specificity of the test?

The company anticipates that of the women who will use the pregnancy-test kit, 10% will actually be pregnant.

***3.97** What is the predictive value positive of the test?

***3.98** Suppose the "cost" of a false negative ($2c$) is twice that of a false positive (c) (because for a false negative prenatal care would be delayed during the first trimester of pregnancy). If the standard home pregnancy-test kit (made by another drug company) has a sensitivity of .98 and a specificity of .98, then which test (the new or standard) will cost the least per woman using it in the general population and by how much?

Mental Health

The Chinese Mini-Mental Status Test (CMMS) consists of 114 items intended to identify people with Alzheimer's disease and senile dementia among people in China [15]. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and nurses and a definitive diagnosis of dementia was made. Table 3.13 shows the results obtained on the subgroup of people with at least some formal education.

Table 3.13 Relationship of clinical dementia to outcome on the Chinese Mini-Mental Status Test

CMMS score	Nondemented	Demented
0–5	0	2
6–10	0	1
11–15	3	4
16–20	9	5
21–25	16	3
26–30	18	1
	46	16

Suppose a cutoff value of ≤ 20 on the test is used to identify people with dementia.

3.99 What is the sensitivity of the test?

3.100 What is the specificity of the test?

3.101 The cutoff value of 20 on the CMMS used to identify people with dementia is arbitrary. Suppose we consider changing the cutoff. What are the sensitivity and specificity if cutoffs of 5, 10, 15, 20, 25, or 30 are used, respectively? Make a table of your results.

3.102 Construct an ROC curve based on the table constructed in Problem 3.101.

3.103 Suppose we want both the sensitivity and specificity to be at least 70%. Use the ROC curve to identify the possible value(s) to use as the cutoff for identifying people with dementia, based on these criteria.

3.104 Calculate the area under the ROC curve. Interpret what this area means in words in the context of this problem.

Demography

A study based on data collected from the Medical Birth Registry of Norway looked at fertility rates according to survival outcomes of previous births [16]. The data are presented in Table 3.14.

3.105 What is the probability of having a livebirth (L) at a second birth given that the outcome of the first pregnancy was a stillbirth (D), that is, death?

3.106 Answer Problem 3.105 if the outcome of the first pregnancy was a livebirth.

3.107 What is the probability of 0, 1, and 2+ additional pregnancies if the first birth was a stillbirth?

3.108 Answer Problem 3.107 if the first birth was a livebirth.

Mental Health

The $\varepsilon 4$ allele of the gene encoding apolipoprotein E (APOE) is strongly associated with Alzheimer's disease, but its value in making the diagnosis remains uncertain. A study was conducted among 2188 patients who were evaluated at autopsy for Alzheimer's disease by previously established pathological criteria [17]. Patients were also evaluated clinically for the presence of Alzheimer's disease. The data in Table 3.15 were presented.

Suppose the pathological diagnosis is considered the gold standard for Alzheimer's disease.

3.109 If the clinical diagnosis is considered a screening test for Alzheimer's disease, then what is the sensitivity of this test?

Table 3.14 Relationship of fertility rates to survival outcome of previous births in Norway

	First birth	Continuing to second birth	Second birth outcome		Continuing to third birth	Third birth outcome	
Perinatal outcome	*n*	*n*		*n*	*n*		*n*
D	7,022	5,924	D	368	277	D	39
						L	238
			L	5,556	3,916	D	115
						L	3,801
L	350,693	265,701	D	3,188	2,444	D	140
						L	2,304
			L	262,513	79,450	D	1,005
						L	78,445

Note: D = dead, L = alive at birth and for at least one week.

Table 3.15 Relationship between clinical and pathological diagnoses of Alzheimer's disease

	Pathological diagnosis	
Clinical diagnosis	Alzheimer's disease	Other causes of dementia
Alzheimer's disease	1643	190
Other causes of dementia	127	228

3.110 What is the specificity of this test?

To possibly improve on the diagnostic accuracy of the clinical diagnosis for Alzheimer's disease, information on both the APOE genotype as well as the clinical diagnosis were considered. The data are presented in Table 3.16.

Suppose we consider the combination of both a clinical diagnosis for Alzheimer's disease *and* the presence of ≥ 1 $\varepsilon 4$ allele as a screening test for Alzheimer's disease.

3.111 What is the sensitivity of this test?

3.112 What is the specificity of this test?

Cardiovascular Disease

A fascinating subject of recent interest is the "Hispanic paradox": Census data "show" that Hispanic people tend to have a lower prevalence of coronary heart disease (CHD) than non-Hispanic whites (NHW) based on health interviews of representative samples of people from different ethnic groups from the U.S. population, although their risk-factor profile is generally worse (more hypertension, diabetes, obesity for Hispanics than for NHW). To study this further, researchers took a group of 1000 Hispanic men ages 50–64 who were free of CHD in 1990 from several counties in Texas and followed them for 5 years. They found that 100 of the men had developed CHD (either fatal cases, or nonfatal cases who survived a heart attack).

3.113 Is the proportion 100 out of 1000 a prevalence rate, an incidence rate, or neither?

Table 3.16 Influence of the APOE genotype in diagnosing Alzheimer's disease (AD)

APOE genotype	Both clinical and pathological criteria for AD	Only clinical criteria for AD	Only pathological criteria for AD	Neither clinical nor pathological criteria for AD
≥ 1 $\varepsilon 4$ allele	1076	66	66	67
No $\varepsilon 4$ allele	567	124	61	161
Total	1643	190	127	228

Given other surveys over the same time period among non-Hispanic whites in these counties, the researchers expected that the comparable rate of CHD for non-Hispanic whites would be 8%.

Another important parameter in the epidemiology of CHD is the *case fatality rate* (the proportion of people who die among those who have a heart attack). Among the 100 CHD cases ascertained among Hispanics, 50 were fatal.

3.114 What is the expected proportion of Hispanic men who will be identified as having a previous heart attack in the past 5 years by health surveys (who are by definition survivors) if we assume the proportion of men with more than one nonfatal heart attack is negligible? What is the comparable proportion for non-Hispanic white men if the expected case-fatality rate is 20% among NHW men with CHD?

3.115 Are these proportions prevalence rates, incidence rates, or neither? Do the results in this problem give insight into why the Hispanic paradox occurs (do Hispanic men truly have lower risk of CHD as government surveys would indicate)? Why or why not?

Genetics

A dominantly inherited genetic disease is identified over several generations of a large family. However, about half of families have dominant disease with *complete penetrance*, whereby if a parent is affected, there is a 50% probability that any one offspring will be affected. Similarly, about half of families have dominant disease with *reduced penetrance*, whereby if a parent is affected, there is a 25% probability that any one offspring will be affected.

Suppose in a particular family one parent and two out of two offspring are affected.

3.116 What is the probability that exactly two out of two offspring will be affected in a family with dominant disease with complete penetrance?

3.117 What is the probability that exactly two out of two offspring will be affected in a family with dominant disease with reduced penetrance?

3.118 What is the probability that the mode of transmission for this particular family is dominant with complete penetrance? Is this a prior probability or a posterior probability?

3.119 Suppose you are a genetic counselor and are asked by the parents what the probability is, if they have another (a third) child, that it will be affected by the disease. What is the answer?

Hypertension

A study was performed in Bergen, Norway [18]. In 1963, 70,000 people in the general population had their blood pressure measured twice. The second reading was used in the analysis. The people were followed for mortality outcome over a 10-year period after the blood-pressure measurement, using death files in the Norwegian Central Bureau of Statistics. Table 3.17 shows results from the subgroup of 5034 men ages 50–59 in 1963:

Table 3.17 Relationship of 10-year mortality to diastolic blood pressure (DBP) at baseline in Norway

	10-year mortality outcome		
DBP (mm Hg)	Dead	Alive	Total
100+	124	295	419
≤99	764	3851	4615
Total	888	4146	5034

3.120 If we regard a diastolic blood pressure of ≥ 100 mm Hg as a screening test for predicting mortality over the next 10 years, then what is the sensitivity of the test?

3.121 What is the specificity of the test?

3.122 If the subjects in the study sample are considered representative of the general population, then what is the predictive value positive and negative of the test?

3.123 Suppose the threshold for positivity were changed from 100+ to 95+. Would the sensitivity and specificity increase, decrease, or remain the same?

SIMULATION—CLASS PROJECT

Infectious Disease

Suppose a standard antibiotic kills a type of bacteria 80% of the time. A new antibiotic is reputed to have better efficacy than the standard antibiotic. Researchers propose to try the new antibiotic on 100 patients infected with the bacteria. Using principles of hypothesis testing (covered in Chapter 7), researchers will deem the new antibiotic "significantly better" than the old if it kills the bacteria for at least 88 out of the 100 infected patients.

3.124 Suppose there is a true probability (true efficacy) of 85% that the new antibiotic will work *for an individual patient.* Perform a "simulation study" on the computer, based on random number generation (using, for example, MINITAB or Excel) for a group of 100 randomly simulated patients. Repeat this exercise 20 times with separate columns for each simulated sample of 100 patients. For what percentage of the 20 samples is the new antibiotic considered "significantly better" than the standard antibiotic? (This percentage is referred to as the *statistical power* of the experiment.) Compare results for different students in the class.

3.125 Repeat the procedure in Problem 3.124, assuming the true efficacy of the new antibiotic is (a), 80%, (b) 90%, (c) 95%, respectively, and compute the statistical power for each of (a), (b), and (c).

3.126 Plot the statistical power versus the true efficacy. Do you think 100 patients is a sufficiently large sample to discover if the new drug is "significantly better" if the true efficacy of the drug is 90%? Why or why not?

Infectious Disease, Cardiovascular Disease

A validation study is to be performed in a local hospital to check the accuracy of assessment of hospital-acquired infection (INF) following coronary bypass surgery (coronary-artery bypass graft, or CABG). In a given year the hospital performs 1100 CABG procedures. A Centers for Disease Control (CDC) algorithm is currently used to categorize subjects as having INF. To validate this algorithm, all CDC^+ subjects ($N = 100$) and a random sample of CDC^- subjects ($N = 1000$) will be ascertained by an infectious-disease (ID) fellow and a detailed investigation will be performed, including a chart review and documentation of antibiotic use. Assume the ID-fellow determination is correct.

Suppose 100 CDC^+ subjects are ascertained, of whom the ID fellow confirms 80. Because there are a large number of CDC^- subjects (1000), only a sample of 100 are studied, of whom the ID fellow confirms 90.

3.127 What is the predictive value positive of the CDC algorithm?

3.128 What is the predictive value negative of the CDC algorithm?

3.129 What is the sensitivity of the CDC algorithm?

3.130 What is the specificity of the CDC algorithm?

Genetics

Suppose a birth defect has a recessive form of inheritance. In a study population, the recessive gene (*a*) initially has a prevalence of 25%. A subject has the birth defect if both maternal and paternal genes are of type *a*.

3.131 In the general population, what is the probability that an individual will have the birth defect, assuming that maternal and paternal genes are inherited independently?

A further study finds that after 10 generations (≈200 years) a lot of inbreeding has taken place in the population. Two subpopulations (populations A and B), consisting of 30% and 70% of the general population, respectively, have formed. Within population A, prevalence of the recessive gene is 40%, whereas in population B it is 10%.

3.132 Suppose that in 25% of marriages both people are from population A, in 65% both are from population B, and in 10% there is one partner from population A and one from population B. What is the probability of a birth defect in the next generation?

3.133 Suppose that a baby is born with a birth defect, but its ancestry is unknown. What is the posterior probability that a baby will have both parents from population A, both parents from population B, or mixed ancestry, respectively? *Hint*: Use Bayes' rule.

Orthopedics

Piriformis syndrome is a pelvic condition that involves malfunction of the piriformis muscle (a deep buttock muscle), which often causes back and buttock pain with sciatica (pain radiating down the leg). An electrophysiologic test to detect piriformis syndrome involves measuring nerve-conduction velocity (NCV) at two nerves in the leg (the tibial and peroneal nerves) with the leg flexed in a specific position. Increases in nerve-conduction velocity in these nerves are often associated with piriformis syndrome. The resulting test, called the FAIR test, is positive if the average NCV in these nerves is delayed by 2+ seconds relative to normal.

A small study compared the FAIR test results with patients' self-reports of how they feel on a visual analog scale (VAS) of 0–10, with 0 indicating no pain and 10 very severe pain. The results were as shown in Table 3.18.

Table 3.18 FAIR test results on piriformis patients

Clinical response	VAS	FAIR ≥ 2	FAIR < 2	Total
Best	≤2	5	14	19
	3–4	3	12	15
	5–6	7	6	13
Worst	≥7	7	6	13
Total		22	38	60

Suppose physicians consider the FAIR test the gold standard, with a FAIR test of ≥2 defined as a true positive and a FAIR test of <2 defined as a true negative. Suppose a VAS of ≤4 is considered a good clinical response based on self-report (a test negative) and a VAS of ≥5 is considered a bad clinical response (a test positive).

3.134 What is the sensitivity of the VAS?

3.135 What is the specificity of the VAS?

3.136 The cutoff points of ≥ 5 for a test positive and ≤ 4 for a test negative are arbitrary. Compute and graph the ROC curve for the VAS test by varying the cutoff point for a test positive. (Use the cutoff points VAS ≥ 0, VAS ≥ 3, VAS ≥ 5, VAS ≥ 7, and VAS ≥ 11 as possible criteria for test positive.)

3.137 The area under the ROC curve is 65%. What does it mean?

Cancer

Breast cancer is considered largely a hormonal disease. An important hormone in breast-cancer research is estradiol. The data in Table 3.19 on serum estradiol were obtained from 213 breast-cancer cases and 432 age-matched controls. All women were age 50–59 years.

Table 3.19 Serum-estradiol data

Serum estradiol (pg/mL)	Cases (N = 213)	Controls (N = 432)
1–4	28	72
5–9	96	233
10–14	53	86
15–19	17	26
20–24	10	6
25–29	3	5
30+	6	4

Suppose a serum-estradiol level of 20+ pg/mL is proposed as a screening criterion for identifying breast-cancer cases.

3.138 What is the sensitivity of this test?

3.139 What is the specificity of this test?

The preceding sample was selected to oversample cases. In the general population, the prevalence of breast cancer is about 2% among women 50 to 59 years old.

3.140 What is the probability of breast cancer among 50- to 59-year-old women in the general population who have a serum-estradiol level of ≥ 20 pg/mL? What is another name for this quantity?

Cardiovascular Disease

Mayo Clinic investigators have tracked coronary-heart-disease (CHD) mortality in Olmstead County, Minnesota, for the past 20 years. Mayo Clinic physicians provided virtually all medical care to Olmstead County residents. Deaths from CHD were subdivided into those that occurred in hospital and out of hospital. In-hospital death rates are thought mainly influenced by advances in medical care. Out-of-hospital death rates are thought mainly influenced by changes in risk-factor levels over time. For men, out-of-hospital CHD death rates were 280 cases per 100,000 men per year and in-hospital CHD death rates were 120 cases per 100,000 men per year in 1998. For women, out-of-hospital CHD death rates were 100 cases per 100,000 women per year; in-hospital CHD death rates were 40 cases per 100,000 women per year in 1998.

3.141 If 50% of the Olmstead County population is male and 50% female, what is the overall CHD mortality rate in Olmstead County in 1998?

The investigators reported that for both men and women, in-hospital CHD death rates were declining at a rate of 5.3% per year, whereas out-of-hospital CHD death rates were declining by 1.8% per year.

3.142 What is the expected overall CHD mortality rate in Olmstead County in 2008 if these trends continue?

3.143 In 2008, what proportion of the CHD deaths will occur in hospital?

3.144 In 2008, what proportion of CHD deaths will be female?

REFERENCES

[1] National Center for Health Statistics. (1976, February 13). *Monthly vital statistics report, advance report, final natality statistics (1974), 24*(11) (Suppl. 2).

[2] Doll, R., Muir, C., & Waterhouse, J. (Eds.). (1970). *Cancer incidence in five continents II.* Berlin: Springer-Verlag.

[3] Feller, W. (1960). *An introduction to probability theory and its applications* (Vol. 1). New York: Wiley.

[4] Podgor, M. J., Leske, M. C., & Ederer, F. (1983). Incidence estimates for lens changes, macular changes, open-angle glaucoma, and diabetic retinopathy. *American Journal of Epidemiology, 118*(2), 206–212.

[5] Punglia, R. S., D'Amico, A. V., Catalona, W. J., Roehl, K. A., & Kuntz, K. M. (2003). Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen. *New England Journal of Medicine, 349*(4), 335–342.

[6] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Diagnostic Radiology, 143,* 29–36.